

# STATISTICAL INFERENCE IN INVERSE PROBLEMS

A Dissertation

by

XIAOLEI XUN

Submitted to the Office of Graduate Studies of  
Texas A&M University  
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2012

Major Subject: Statistics

# STATISTICAL INFERENCE IN INVERSE PROBLEMS

A Dissertation

by

XIAOLEI XUN

Submitted to the Office of Graduate Studies of  
Texas A&M University  
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Co-Chairs of Committee,	Raymond J. Carroll
	Bani K. Mallick
Committee Members,	Peter Kuchment
	Huiyan Sang
Head of Department,	Simon Sheather

May 2012

Major Subject: Statistics

## ABSTRACT

Statistical Inference in Inverse Problems. (May 2012)

Xiaolei Xun, B.S., Zhejiang University;

M.S., Texas A&M University

Co-Chairs of Advisory Committee: Dr. Raymond J. Carroll  
Dr. Bani K. Mallick

Inverse problems have gained popularity in statistical research recently. This dissertation consists of two statistical inverse problems: a Bayesian approach to detection of small low emission sources on a large random background, and parameter estimation methods for partial differential equation (PDE) models.

Source detection problem arises, for instance, in some homeland security applications. We address the problem of detecting presence and location of a small low emission source inside an object, when the background noise dominates. The goal is to reach the signal-to-noise ratio levels on the order of  $10^{-3}$ . We develop a Bayesian approach to this problem in two-dimension. The method allows inference not only about the existence of the source, but also about its location. We derive Bayes factors for model selection and estimation of location based on Markov chain Monte Carlo simulation. A simulation study shows that with sufficiently high total emission level, our method can effectively locate the source.

Differential equation (DE) models are widely used to model dynamic processes in many fields. The forward problem of solving equations for given parameters that define the DEs has been extensively studied in the past. However, the inverse problem of estimating parameters based on observed state variables is relatively sparse in the statistical literature, and this is especially the case for PDE models. We propose two joint modeling schemes to solve for constant parameters in PDEs: a parameter cascading method and a Bayesian treatment. In both methods, the unknown functions

are expressed via basis function expansion. For the parameter cascading method, we develop the algorithm to estimate the parameters and derive a sandwich estimator of the covariance matrix. For the Bayesian method, we develop the joint model for data and the PDE, and describe how the Markov chain Monte Carlo technique is employed to make posterior inference. A straightforward two-stage method is to first fit the data and then to estimate parameters by the least square principle. The three approaches are illustrated using simulated examples and compared via simulation studies. Simulation results show that the proposed methods outperform the two-stage method.

To my family

## ACKNOWLEDGMENTS

First of all, I would like to thank my committee co-chairs, Dr. Raymond J. Carroll and Dr. Bani K. Mallick, and my committee members, Dr. Peter Kuchment and Dr. Huiyan Sang, for their guidance and support throughout the course of the research. I am fortunate enough to have them to guide my work at Texas A&M University.

I also want to thank Dr. Michael Longnecker for his kindness and helpfulness, especially in teaching. Thanks also go to all my friends and colleagues, the department faculty and staff members, who make my time at Texas A&M University a great experience. I am very grateful for having you around in my effort toward my Ph.D. in Statistics.

## TABLE OF CONTENTS

CHAPTER		Page
I	INTRODUCTION . . . . .	1
II	A BAYESIAN APPROACH TO DETECTION OF SMALL LOW EMISSION SOURCES . . . . .	5
	2.1. Introduction . . . . .	5
	2.2. Models . . . . .	8
	2.2.1. The Model without a Source . . . . .	9
	2.2.2. The Model with a Source . . . . .	9
	2.3. Model Selection via Bayes Factors . . . . .	11
	2.3.1. Computation . . . . .	13
	2.3.2. Algorithm Summary . . . . .	16
	2.4. Simulation Study . . . . .	17
	2.5. Concluding Remarks . . . . .	21
III	PARAMETER ESTIMATION OF PARTIAL DIFFEREN- TIAL EQUATION MODELS . . . . .	23
	3.1. Introduction . . . . .	23
	3.2. Basis Function Approximation . . . . .	27
	3.2.1. B-Spline Basics . . . . .	30
	3.3. Parameter Cascading Method . . . . .	30
	3.3.1. Estimating PDE Parameters . . . . .	31
	3.3.2. Smoothing Parameter Selection . . . . .	34
	3.3.3. Variance Estimation of Parameters . . . . .	34
	3.4. Bayesian Estimation and Inference . . . . .	36
	3.4.1. Bayesian P-Splines . . . . .	39
	3.5. Simulations . . . . .	39
	3.5.1. A Two-Stage Method . . . . .	40
	3.5.2. Data Generating Mechanism . . . . .	40
	3.5.3. Performance of the Proposed Methods . . . . .	43
	3.6. Empirical Example . . . . .	48
	3.6.1. The Example . . . . .	48
	3.6.2. Results . . . . .	48
	3.7. Concluding Remarks . . . . .	54

CHAPTER	Page
IV CONCLUSION . . . . .	55
REFERENCES . . . . .	57
APPENDIX A . . . . .	63
VITA . . . . .	75



## LIST OF TABLES

TABLE		Page
1	Summary of Bayes factors for simulation in Section 2.4. Sample size $n = 2 \times 10^5$ . There are 10 simulations performed at each combination of level $p$ and location, and 20 simulations at $p = 0$ . The values reported in the table are minimum, median, maximum of 10 Bayes factors, and the proportion of Bayes factors being greater than 3. In the last column is the median of $\text{pr}(p = 0 \tilde{Y})$ calculated from the 10 data sets. . . . .	19
2	Repeat of the Table 1 with $5 \times 10^5$ samples. . . . .	19
3	The biases, standard deviations (SDs), square root of mean squared errors (RMSEs) of the parameter estimates for the PDE model (3.2) using the Bayesian method (BM), the parameter cascading method (PC), and the two-stage method (TS) in the 1000 simulation replicates. The coverage probabilities (CP) of on 95% credible/confidence intervals are also shown. The true parameter values are shown in the second row. As the two-stage method results in significant bias, we skip variance calculation for this method, and no coverage probability is provided. . . . .	44
4	Results of the empirical example. We show the estimates for each parameter, and corresponding standard errors (SE). Methods: BM = Bayesian method; PC = parameter cascading method; TS = two-stage method. . . . .	50
5	Summary of function estimation in the empirical example. Estimated square root of average squared errors of $g(t, z)$ and PDE $\mathcal{F}(\cdot)$ estimation, $\widehat{\text{RSAE}}(\hat{g})$ and $\widehat{\text{RSAE}}(\hat{\mathcal{F}})$ , are shown. Methods: BM = Bayesian method; PC = parameter cascading method; TS = two-stage method. . . . .	50

## LIST OF FIGURES

FIGURE		Page
1	Experiment set-up. A direction sensitive (e.g., collimated or Compton type) detector determines the normal parameters $(\theta, S)$ of the trajectory of the incoming particle. The detected particles might be either emitted from the source or coming from random background.	7
2	Snapshots of $f(\ell_1, \ell_2 p, \tilde{Y})$ . The data set contains 200,000 samples. The true emission rate is $p = 0.001$ , and the source is located at $(0.3, 0.6)$ . The left figure is conditioned at $p = 0.0002$ ; the right one assumes $p = 0.001$ . The multimodality illustrates the difficulty of MCMC sampling in this problem. . . . .	13
3	The estimated location of the source with $n = 5 \times 10^5$ sample counts for various emitting levels, with 95% highest posterior density region. The top left plot is with $p = 0.01$ ; the top right plot is with $p = 0.005$ ; the bottom left plot is with $p = 0.001$ ; the bottom right plot is with $p = 0$ . In the plots where a source exists, the location of its center is indicated by the intersection of gray dashed lines. Each of the above figures is plotted with the posterior sample having median Bayes factors among the 10 simulated cases. . . . .	20
4	Example of building block of B-splines. This is a 2D quartic B-spline basis function formed by tensor product of 1D quartic B-spline functions, at knot $(0.5, 0.5)$ . . . . .	31
5	Snapshots of the solution function $g(t, z)$ , i.e., the error-free data. Top: 3-D plot of the surface $g(t, z)$ . Middle: plot of $g(t_i, z)$ for time values $t_i$ over range, with indices $i = 1, 6, 11, 16, 20$ . Bottom: plot of $g(t, z_j)$ for range values $z_j$ over time, with indices $j = 1, 11, 21, 31, 40$ . . . . .	41

## FIGURE

## Page

6	Snapshots of the partial derivatives of $g(t, z)$ . Top: plot of $\partial g(t, z_j)/\partial t$ for range values $z_j$ over time, with indices $j = 1, 11, 21, 31, 40$ . Middle: plot of $\partial g(t_i, z)/\partial z$ for time values $t_i$ over range with indices $i = 1, 6, 11, 16, 20$ . Bottom: plot of $\partial^2 g(t_i, z)/\partial z^2$ for time values $t_i$ over range, with indices $i = 1, 6, 11, 16, 20$ . . . . .	42
7	Boxplots of the square root of average squared errors (RASE) from 1000 data sets in the simulation study. Left: boxplots of $\text{RASE}(\hat{g})$ , the solution function estimation, by all three methods. Right: boxplots of $\text{RASE}(\hat{\mathcal{F}})$ , the PDE model estimation, by all three methods. The three methods produce similar data fitting, but the parameter cascading and Bayesian methods result in better PDE fitting. . . . .	45
8	Cross-sectional views of the estimated solution in one data set in the simulation study. Left: function $\hat{g}(t_{11}, z)$ . Right: function $\hat{g}(t, z_{20})$ . All three methods produce similar function estimation. . . .	45
9	Estimated partial derivatives of the curves shown in Figure 8. Top: function $\partial \hat{g}(t_{11}, z)/\partial z$ . Middle: function $\partial \hat{g}(t, z_{20})/\partial t$ . Bottom: function $\partial^2 \hat{g}(t_{11}, z)/\partial z^2$ . There is a clear difference in the estimated second partial derivative between the two-stage method and the joint modeling methods. . . . .	47
10	Snapshots of empirical data. Top: 3D plot of the received signal. Middle: the received signal at a few burst values over range. Bottom: the received signal at a few range values over burst. . . . .	49
11	Cross-sectional views of estimated solution in the empirical example. Top: function $\hat{g}(t_{11}, z)$ . Bottom: function $\hat{g}(t, z_{30})$ . Three methods produce similar function estimation. . . . .	52
12	Cross-sectional views of estimated derivatives of curves shown in Figure 11. Top: function $\partial \hat{g}(t_{11}, z)/\partial z$ . Middle: function $\partial \hat{g}(t, z_{30})/\partial t$ . Bottom: function $\partial^2 \hat{g}(t_{11}, z)/\partial z^2$ . . . . .	53

## CHAPTER I

### INTRODUCTION

Inverse problems arise widely in many different fields, from science like physics and biology to engineering like medical imaging and remote sensing. Generally speaking, inverse problems are concerned with converting observed measurements into information about quantities of a physical system, which are of primary interest but could not be observed directly. This sounds similar to traditional statistical problems, but they have very different taste. Inverse problems usually involve a model, which is governed by physical principle and describes the underlying process, instead of using an arbitrary model to fit the data. For example, a challenging inverse problem occurs when we try to measure the temperature inside a furnace. The ubiquitous thermometer containing quicksilver could not be used because of the extremely high temperature. An alternative technique is to use ultrasound, as acoustic properties of the gas inside the furnace are changed by the heat. The physical model in this problem describes the acoustic wave propagation, with temperature as (functional) parameter. And the inverse problem in this situation is to estimate the temperature from ultrasound observations.

Inverse problems, as mathematical problems, have a long history and are well studied. However, they are relatively new in the statistical community, but gaining in popularity in recent years. From a mathematical point of view, we invert some operator in an inverse problem on a basis of deterministic models. In general, the problem could be stated as follows: let  $F$  be an operator mapping from model parameter space  $M$  to data space  $D$ , the inverse problem is then to find the parameter  $m \in M$ ,

---

This dissertation follows the style of *Biometrics*.

given observed data  $y \in D$ , such that  $y = F(m)$  is satisfied or best approximated. A well-posed problem has unique and stable solution, which depends on the data continuously. Many inverse problems are not well-posed, in other words, ill-posed. The issue is usually on stability. Regularization techniques are used to treat ill-posed inverse problems. See Kaipio and Somersalo (2005) for a survey of most commonly used methods, including truncated singular value decomposition, Tikhonov regularization and several truncated iterative methods.

From a statistical point of view, inverse problems are recast as problems of statistical inference, such as parameter estimation, probability density estimation, model selection problem, etc. The objective is to extract information and quantify the uncertainty in the inference procedure. Bayesian statistics is widely used in this area. Following the Bayes' theorem, problems are formulated under a unified framework and solved in a systematical way. As always, all unknown quantities are modeled as random variables, prior distributions are assigned based on information available before measurements, and the posterior distribution is the Bayesian solution. Challenges arise in the construction of prior distributions and likelihood functions, especially the priors, which must be handled with great care.

Frequentist methods are also applicable to inverse problems. But unlike Bayesian approaches, solutions are problem specific. Also, incorporating all available information into model could be challenging for frequentists, whereas Bayesian methodology has the advantage of achieving this goal naturally.

This dissertation consists of two statistical inverse problems: a Bayesian approach to detection of small low emission sources on a large random background, and parameter estimation methods for partial differential equation models. Brief introduction to these topics is given in the following paragraphs.

Source detection problem arises, for instance, in some homeland security appli-

cations. We consider the problem of detecting existence of a low emission radiating source inside a volume, in the presence of a strong random background. We are interested in the situation when about 99.9% of the total detections come from the background particles and from the particles emitted by the source that have been scattered. In other words, only about 1% of detected hits are by the ballistic particles coming from the source. Although there is probably no general solution, in the applications we have in mind, the radiating source, if present, would be significantly smaller than the whole object. The availability of detectors determining direction of an incoming particle makes detection conceivable. The idea is that if a source is present, ballistic particles coming from it might lead to a statistically significant increase in the number of trajectories through the source, and thus to detection. Under appropriate conditions, this happens to be the case. See the discussion in Allmaras et al. (2010) for details. We develop statistical models for each of the two cases, existing source or non-existing source, and then decide, based on the collected data and the value of the corresponding Bayes factor, which model fits better the collected data. A simulation study shows that with sufficiently high total emission level, our method can effectively locate the source.

Differential equation (DE) models are widely used to model dynamic processes in many fields, for example, engineering and biomedical sciences. The forward problem of solving equations or simulating state variables for given parameters that define the DE models has been extensively studied in the past. However, the inverse problem of estimating parameters based on observed state variables is relatively sparse in the statistical literature, and this is especially the case for partial differential equations (PDE). We propose two methods to solve for parameters in PDE models: a Bayesian treatment and a parameter cascading method. In both methods, the unknown function is expressed via basis function expansion. For the Bayesian method, we develop

the joint model for data and PDE, and describe the Markov chain Monte Carlo (MCMC) technique to generate observations from the posterior distribution. For the parameter cascading method, we develop the algorithm to estimate the parameter and derive a sandwich estimator of the covariance matrix. As a strawman, we also extend a straightforward two-stage method for ordinary differential equations, which first fits the data and then estimates parameters by least square principle, to PDE models. The approaches are illustrated using simulated examples and compared via simulation studies. Simulation studies show that either Bayesian method or parameter cascading method are more statistical efficient than the two-stage method.

## CHAPTER II

### A BAYESIAN APPROACH TO DETECTION OF SMALL LOW EMISSION SOURCES\*

#### 2.1. Introduction

We consider the problem of detecting existence of a low emission radiating source inside a volume, in the presence of a strong random background. One can easily imagine possible applications of such detection, for instance to homeland security. We are interested in the situation when about 99.9% of the total detections come from the background particles and from the particles emitted by the source that have been scattered (we will consider the latter as a part of the background). In other words, only about 1% of detected hits are by the ballistic particles coming from the source. Most of the background particles are not emitted inside the object, but rather are present in the surrounding environment (e.g., cosmic rays). Although medical emission tomographic imaging faces similar problems (e.g., Budinger, Gullberg and Huesman, 1979), the overwhelming level of noise that has just been mentioned would be considered impossible to handle there. So, how can one attack this problem? Although there is probably no general solution, in the applications we have in mind, the radiating source, if present, would be significantly (on the order of hundred times) geometrically smaller than the whole object. As is explained in Allmaras et al. (2010), this, and the availability of detectors determining direction of an incoming particle, make detection conceivable under appropriate conditions. In this text, we consider

---

\* Reprinted with permission from “A Bayesian approach to the detection of small low emission sources” by Xun, X., Mallick, B. K., Carroll, R. J., and Kuchment, P., 2011. *Inverse Problems*, 27, 155009, Copyright 2011 by IOP Publishing Ltd. Online version DOI: 10.1088/0266-5611/27/11/115009.



the  $2D$  problem. Unlike Allmaras et al. (2010), where more analytic techniques are considered, we propose a Bayesian method, which allows inference about the existence and location of the possible source.

The problem can be stated as follows. One is interested in certain type of particles (say,  $\gamma$ -photons or neutrons, although the type of particles is irrelevant for our purpose). Suppose that the observed area belongs to the unit disk  $D$  in the  $l_1$ - $l_2$ -plane (see Figure 1). Detectors, placed around the object, are assumed to be able to determine the linear trajectory of each incoming particle. It is assumed that detectors surround the object in such a way that any escaping particle hits a detector (this assumption could in principle be weakened). Most of (or maybe all) detected particles are coming from a random background (and, in particular, are not emitted inside the object). Besides the background emission, a small (in comparison with the total object's size) source might be present, whose emission is assumed to be very low in comparison with the background. Many of the particles emitted by the source will be scattered, and only a small number of them will reach the detectors unscattered (ballistic). The goal is to detect the presence of such an object, if the emission is dominated by the background, e.g., such that the ballistic particles coming from a possible source could account to about 0.1% of the total emission. In this initial study, the effect of scattering of particles emitted by the source is neglected. Due to the negligible size of the scattered emission with respect to the background, this should not be a serious restriction.

The set-up is illustrated in Figure 1. The trajectory of a particle that hits the detector can be identified by its normal coordinates  $(\theta, S)$ , and thus we assume that the detector provides the values  $(\theta, S)$  for each hit.

We expect that a radiation source of a small radius  $d$  might be present, in which case we denote its location point as  $L = (\ell_1, \ell_2)$ . If a particle is emitted from this

source and reaches the detectors ballistically (unscattered), then  $\theta$  and  $S$  satisfy the inequality

$$|\ell_1 \cos(\theta) + \ell_2 \sin(\theta) - S| \leq d. \quad (2.1)$$

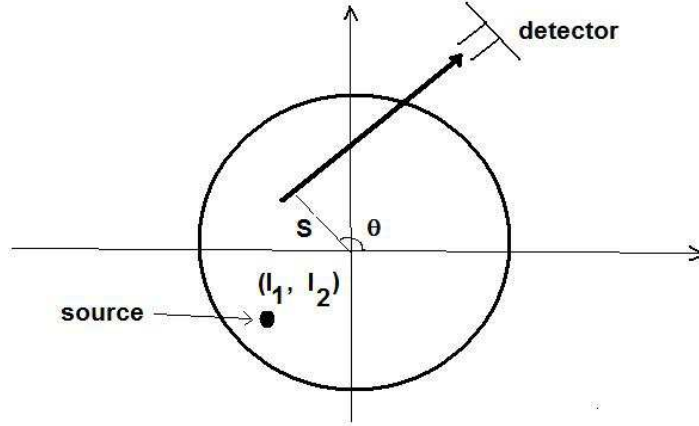


Figure 1. Experiment set-up. A direction sensitive (e.g., collimated or Compton type) detector determines the normal parameters  $(\theta, S)$  of the trajectory of the incoming particle. The detected particles might be either emitted from the source or coming from random background.

Most particles from the random background normally will not satisfy this condition, but a small portion might. The idea is that if a source is present, ballistic particles coming from it might lead to a statistically significant increase in the number of trajectories satisfying (2.1), and thus to detection. Under appropriate conditions (geometrically sufficiently small source and sufficiently large total count in the sample), this happens to be the case (see the discussion in Allmaras et al., 2010).

In the rest of this chapter, we first introduce the candidate models describing the situations without a source and with a source, respectively. Then we explain the

calculation of Bayes factors for determining the presence of a source, as well as the computational details of our Markov chain Monte Carlo (MCMC) algorithm. Finally we examine the performance of the method for various levels of source emission rate via simulation studies. The simulation results confirm the possibility of detection.

## 2.2. Models

Suppose that the direction sensitive detectors registered hits by  $n$  particles and recorded the corresponding normal coordinates  $(\theta_i, S_i)$  for  $i = 1, \dots, n$  of their incoming directions. We denote by  $\delta_i$  the (unobserved) indicator that the  $i^{th}$  particle is coming ballistically from the suspected source. In other words,  $\delta_i = 1$  if the  $i^{th}$  particle comes from the source. Otherwise,  $\delta_i = 0$ .

If there is no source present, then  $\text{pr}(\delta_i = 1) = 0$ . If there is a source, we assume that the  $\delta_i$ 's are independently and identically distributed according to the Bernoulli( $p$ ) law. This covers also the possible absence of a source, in which case  $p = 0$ .

Our plan is to develop statistical models for each of these cases, and then decide, based on the collected data and the value of the corresponding Bayes factor, which model fits better the collected data.

We would like to point out that this is not expected to be a simple problem, since it involves inference as to whether a non-negative parameter  $p$  takes its boundary value  $p = 0$ . Even in simple variance components models, frequentist boundary value testing is a difficult matter, see for example Crainiceanu and Ruppert (2004).

### 2.2.1. The Model without a Source

When there is no source (we call this model  $M_1$ ), all hits at the detectors come from the random background and thus  $\delta_i = 0$  for all  $i = 1, \dots, n$ . We will assume in this text that the random background is isotropic and uniform. In other words, the angle  $\theta$  and the distance  $S$  from the origin of the trajectory are uniformly distributed:

$$[\theta_i | \delta_i = 0] = \text{Uniform}(0, 2\pi); \quad (2.2)$$

$$[S_i | \delta_i = 0] = \text{Uniform}(-1, 1). \quad (2.3)$$

Notice that particles having trajectories with  $|S| > 1$  do not get detected and thus do not enter the model.

### 2.2.2. The Model with a Source

If a source exists (model  $M_2$ ), then  $\text{pr}(\delta_i = 1) = p > 0$ . If the particle comes from the background, then  $\delta_i = 0$  and relations (2.2) still hold. Assuming that the source in question is isotropic and uniform, when  $\delta_i = 1$ , we have the following distributions:

$$[\theta_i | \delta_i = 1] = \text{Uniform}(0, 2\pi); \quad (2.4)$$

$$[S_i | \theta_i, (\ell_1, \ell_2), \delta_i = 1] = \ell_1 \cos(\theta_i) + \ell_2 \sin(\theta_i) + \text{Uniform}(-d, d). \quad (2.5)$$

Our goal thus is to choose between the models  $M_1$  and  $M_2$ , based on the measured data. We show in the following subsection the priors, likelihood and the posterior distribution associated with model  $M_2$ .

### 2.2.2.1. Likelihood and Posterior of Model with a Source

For the model  $M_2$  with a source, see (2.5), parameters of interest are  $\phi = (p, \ell_1, \ell_2)$ .

Priors are assigned as follows:

$$L = (\ell_1, \ell_2) = \text{Uniform}\{\ell_1^2 + \ell_2^2 \leq 1\},$$

i.e., *a priori* the source can be located anywhere with equal probability. Also, we allow  $p$  to have the uniform discrete distribution on the set  $\{0 < p_1, \dots, p_h\}$ , which contains  $h$  values equally spaced on the interval  $[p_1, p_h]$ . This set is chosen based on *a priori* estimates of the possible emission strength of the source.

For the implementation of the MCMC algorithm, it is convenient (see Section 2.3.1.2) to use the polar coordinates  $(r, u)$  of the center  $L$ :

$$\ell_1 = r \cos u, \quad \ell_2 = r \sin u.$$

Denoting the new parameterization by  $\psi = (p, r, u)$ , the prior  $f(\psi|M_2)$  is

$$f(\psi|M_2) = f(p, r, u|M_2) = h^{-1}\pi^{-1}r\text{I}(0 \leq r \leq 1)\text{I}(0 \leq u \leq 2\pi),$$

where  $\text{I}(\cdot)$  is the indicator function.

Let  $Y_i = (\theta_i, S_i)$  denote the  $i^{\text{th}}$  observation (i.e., the normal coordinates of the trajectory (at the detector) of the  $i^{\text{th}}$  detected particle) and, as before, let  $\delta_i$  be the (unobserved) indicator associated with it. It will be convenient to introduce vectors  $\tilde{Y} = (Y_1, \dots, Y_n)$  and  $\tilde{\delta} = (\delta_1, \dots, \delta_n)$ . Then the likelihood function is

$$\begin{aligned} & f(\tilde{Y}|p, r, u, M_2) \\ &= \prod_{i=1}^n (4\pi)^{-1} [pd^{-1}\text{I}\{|r \cos(u) \cos(\theta_i) + r \sin(u) \sin(\theta_i) - S_i| \leq d\} + 1 - p] \\ &= (4\pi)^{-n} \prod_{i=1}^n [pd^{-1}\text{I}\{|r \cos(u - \theta_i) - S_i| \leq d\} + 1 - p] \end{aligned} \tag{2.6}$$

$$= (4\pi)^{-n}(1-p)^{n-J}(pd^{-1}+1-p)^J,$$

where  $J = \sum_{i=1}^n \mathbf{I}\{|r \cos(u - \theta_i) - S_i| \leq d\}$  counts the total number of particles whose incoming trajectories at the detectors pass near the location  $(r, u)$ .

Given the above prior and likelihood, the posterior is

$$f(\psi|\tilde{Y}, M_2) = f(p, r, u|\tilde{Y}, M_2) \propto r(1-p)^{n-J}(pd^{-1}+1-p)^J, \quad (2.7)$$

where  $p \in \{p_1, \dots, p_h\}$ ,  $r \in [0, 1]$  and  $u \in [0, 2\pi]$ .

### 2.3. Model Selection via Bayes Factors

Let  $\text{pr}(M_j)$  be the prior probability of model  $M_j$  and  $\text{pr}(\tilde{Y}|M_j)$  be the marginal distribution of the data, given model  $M_j$ , where  $j = 1, 2$ . We also denote by  $\text{pr}(M_j|\tilde{Y})$  the posterior probability of the model  $M_j$ .

The parameters of interest are  $p$  and  $L = (\ell_1, \ell_2)$ . Indeed, if  $p = 0$ , then there is no source, while if  $p > 0$ , the source is present at the location  $L$ .

In the next section, we describe the Bayesian approach that will be used for model selection. Then the computation and algorithm will be explained.

We use a Bayes factor approach to select between the two models  $M_1$  and  $M_2$  in question. The Bayes factor is defined as the ratio of the prior and posterior odds:

$$\begin{aligned} BF &= \frac{\text{pr}(M_1)/\text{pr}(M_2)}{\text{pr}(M_1|\tilde{Y})/\text{pr}(M_2|\tilde{Y})} \\ &= \frac{\text{pr}(M_1)\text{pr}(M_2)\text{pr}(\tilde{Y}|M_2)}{\text{pr}(M_2)\text{pr}(M_1)\text{pr}(\tilde{Y}|M_1)} = \frac{\text{pr}(\tilde{Y}|M_2)}{\text{pr}(\tilde{Y}|M_1)}. \end{aligned}$$

This number serves as an indicator of which of the models  $M_1$  and  $M_2$  is more supported by the data. If  $BF > 1$ , this indicates  $M_2$  being more strongly supported by the data. Otherwise,  $M_1$  is more strongly supported. Furthermore, the magnitude

of the Bayes factor is a measure of how strong the evidence is for or against  $M_1$ . According to Kass and Raftery (1995), when the Bayes factor exceeds 3, 20 and 150, one can say that, correspondingly, a positive, strong, and overwhelming evidence exists that a source is present. See Jeffreys (1961), Evett (1991) and Good (1985) for further interpretation of Bayes factors.

We thus need the marginal distributions  $\text{pr}(\tilde{Y}|M_j)$  to be calculated for each candidate model  $M_j, j = 1, 2$ .

Under the null model  $M_1$ , in which there is no source, one concludes that the corresponding marginal probability density of  $\tilde{Y}$  is:

$$\text{pr}(\tilde{Y}|M_1) = (4\pi)^{-n}.$$

When there is a source, we denote by  $\Psi$  the **sample space of parameters under  $M_2$** . The points  $\psi$  from this space is the triples  $\psi = (p, r, u)$  described in Section 2.2.2.1. Then the marginal probability of  $\tilde{Y}$  under  $M_2$  is

$$\text{pr}(\tilde{Y}|M_2) = \int_{\Psi} \text{pr}(\tilde{Y}, \psi|M_2) f(\psi|M_2) d\psi,$$

a quantity that cannot be computed explicitly.

The usual Monte Carlo method of computation is as follows. Suppose we have  $k = 1, \dots, K$  samples  $\psi^{(k)} = (p^{(k)}, r^{(k)}, u^{(k)})$  from the posterior distribution. The marginal distribution  $\text{pr}(\tilde{Y}|M_2)$  can then be estimated as

$$\widehat{\text{pr}}(\tilde{Y}|M_2) \approx \{K^{-1} \sum_{k=1}^K \text{pr}(\tilde{Y}|\psi^{(k)}, M_2)^{-1}\}^{-1},$$

i.e., **the harmonic mean of the likelihoods**  $\text{pr}(\tilde{Y}|\psi, M_2)$  (see, e.g., Kass and Raftery, 1995). Given this, the Bayes factor is calculated as

$$\widehat{\text{BF}} = \widehat{\text{pr}}(\tilde{Y}|M_2)/\text{pr}(\tilde{Y}|M_1) \quad (2.8)$$

$$= \left[ K^{-1} \sum_{k=1}^K \{ (1 - p^{(k)})^{n-J(k)} (p^{(k)} d^{-1} + 1 - p^{(k)})^{J(k)} \}^{-1} \right]^{-1},$$

where  $J(k) = \sum_{i=1}^n \mathbf{I}(|r^{(k)} \cos(u^{(k)} - \theta_i) - S_i| \leq d)$ . See Kass and Raftery (1995) and Raftery (1996) for more details about calculation of Bayes factors.

Computational details of the MCMC implementations and calculation of Bayes factors are shown in the following subsections.

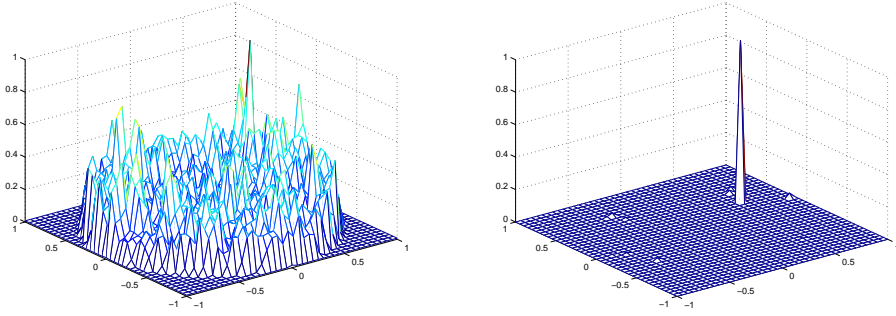


Figure 2. Snapshots of  $f(\ell_1, \ell_2 | p, \tilde{Y})$ . The data set contains 200,000 samples. The true emission rate is  $p = 0.001$ , and the source is located at  $(0.3, 0.6)$ . The left figure is conditioned at  $p = 0.0002$ ; the right one assumes  $p = 0.001$ . The multimodality illustrates the difficulty of MCMC sampling in this problem.

### 2.3.1. Computation

With the model and prior in Section 2.2.2, the posterior distribution is not straightforward to sample from. Thus, the Markov chain Monte Carlo method is used to simulate the parameters from the posterior.

Standard implementation of the Gibbs sampler in this problem will not work, since we discover, as Figure 2 shows, that the posterior distributions are extremely multimodal. The reason for this multimodality is clear. Indeed, what the algorithm essentially does is to look at concentrations of trajectories at different locations. If



the threshold is set too low, as in the left part of Figure 2, one expects to find (and indeed finds) such concentrations in quite a few places.

To overcome this problem, after some experimentation we adopted a parallel tempering method in order to improve mixing of simulations from this multimodal distribution. We describe now the algorithm to sample from  $f(\psi|\tilde{Y}, M_2)$  in more details. The reader interested only in the results of the implementation, can skip the following sub-sections and move directly to Section 2.4.

### 2.3.1.1. Implementing the Parallel Tempering Algorithm

One can find discussions of parallel tempering algorithms in Goswami and Liu (2007) and Liang, Liu and Carroll (2010).

We run  $N$  parallel chains, each with equilibrium  $f_i(x) \propto f(x)^{1/T_i}$ , where  $f(x)$  is the target posterior distribution  $f(p, r, u|\tilde{Y}, M_2)$  and  $T_i$  is a given temperature level. The temperature ladder  $T_1 > \dots > T_N = 1$  plays the most important role in the algorithm, and is constructed in the following (trial-and-error) manner. We decide first the highest temperature such that a single MCMC run (e.g., using Metropolis-Hastings algorithm) at that temperature can explore the whole sample space easily (e.g., the acceptance rate of MH is about 90%). Then the next temperature level is chosen such that the rate of exchanging samples with the chain at previous temperature is moderate (e.g., 20%). We have found in numerical experiments that  $N = 6$  works well, with highest temperature being 5 and exponentially decreasing to 1. In our MCMC simulation, the Gibbs algorithm is implemented at each chain, and all chains start with random values. Let  $\tilde{x}^{(t)} = (x_1^{(t)}, \dots, x_N^{(t)})$  denote the current population of samples from  $N$  chains.

During each iteration of MCMC, the following steps of mutation and exchange are implemented.

Mutation: Update  $x_i^{(t)}$  to  $x_i^{(t+1)}$  by the Gibbs sampler, for  $i = 1, \dots, N$ . Details are shown in the subsection 2.3.1.2.

Exchange: Starting from the first chain, try to swap with neighbors as follows,

- For  $i = 1, N$ , exchange with the only neighbor; for  $i = 2, \dots, N - 1$ , exchange with the two neighbors with equal probability.
- Then accept the exchange of states  $i$  and  $j$  with probability

$$\min \left\{ 1, \exp \left( \left[ \log f(x_j^{(t+1)}) - \log f(x_i^{(t+1)}) \right] \cdot [T_i^{-1} - T_j^{-1}] \right) \right\}.$$

The chain with  $T_N = 1$ , which has the target posterior distribution as equilibrium, is used in the harmonic mean estimate of Bayes factors.

### 2.3.1.2. Implementing the Gibbs Sampler

In each Gibbs update, the target distribution is one of the  $f_i$ 's. In the following context, a function  $h(x|\cdot)$  refers to the full conditional distribution of  $X$ , given all the other unknown variables. Notice that the unknown additive constants in the logarithm of a distribution do not affect Gibbs sampler.

Joint Distribution in Each Iteration: It is easily seen that the joint posterior distribution of  $(p, r, u)$  is given as

$$\begin{aligned} \log\{f_i(p, r, u|\tilde{Y})\} &= T_i^{-1} \log\{f(p, r, u|\tilde{Y})\} + C \\ &= T_i^{-1} [\log(r) + \log\{f(\tilde{Y}|p, r, u)\}] + C. \end{aligned}$$

Updating  $p$ : It is easily seen that

$$\log\{f_i(p|\cdot)\} = T_i^{-1} \{(n - J) \log(1 - p) + J \log(pd^{-1} + 1 - p)\} + C_p.$$

To update  $p$  for each  $f_i(p|\cdot)$ , the following steps are taken.

- Compute the full conditional distribution  $f_i(p_j|\cdot) = \pi_j$ , for  $j = 1, \dots, h$ .
- Form  $\omega_j = \pi_j / \sum_{k=1}^h \pi_k$ , for  $j = 1, \dots, h$ .
- Draw from the vector  $(p_1, \dots, p_h)$  with probabilities  $(\omega_1, \dots, \omega_h)$ .

Updating the Radius Component  $r$  in the Polar Coordinates: It is easily seen that

$$\log\{f_i(r|\cdot)\} = T_i^{-1}\{J\log(pd^{-1} + 1 - p) - J\log(1 - p) + \log(r)\} + C_r.$$

To update  $r$ , the Metropolis-Hastings algorithm is implemented. The proposal is a normal distribution  $N(r_{\text{curr}}, \sigma_{\text{prop},r})$  truncated on the interval  $[0, 1]$ , where the mean  $r_{\text{curr}}$  is the current value and the standard deviation  $\sigma_{\text{prop},r}$  is a constant.

Updating the Angle Component  $u$  in the Polar Coordinates: It is easily seen that

$$\log\{f_i(u|\cdot)\} = T_i^{-1}J\{\log(pd^{-1} + 1 - p) - \log(1 - p)\} + C_u.$$

To update  $u$ , again the Metropolis-Hastings algorithm is implemented. Notice that the target function is periodic in  $u$ , and  $u$  is restricted to  $[0, 2\pi]$ . To ease movement across boundary, we propose a value  $u_{\text{prop}}$  from the normal distribution  $N(u_{\text{curr}}, \sigma_{\text{prop},u}^2)$ , where the mean  $u_{\text{curr}}$  is the current value and standard deviation  $\sigma_{\text{prop},u}$  is a constant. If  $u_{\text{prop}} > 2\pi$  or  $u_{\text{prop}} < 0$ , then the candidate is reset to be  $u_{\text{prop,actual}} = u_{\text{prop}} \bmod(2\pi)$ .

### 2.3.2. Algorithm Summary

Our Bayesian approach could be summarized as follows. Given a particular dataset  $\tilde{Y}$ , denote the posterior by  $f(\psi|\tilde{Y}, M_2)$  as in equation (2.7),

1. decide the hyperparameters  $a_p$  and  $b_p$  ( $b_p > a_p > 0$ ) based on prior knowledge of  $p$ , and the grid  $a_p = p_1 < \dots < p_h = b_p$  according to desired precision in  $p$ ,

2. decide the temperature ladder  $T_1 > \cdots > T_N = 1$  and define, for  $i = 1, \dots, N$ ,  

$$f_i(\psi) = \{f(\psi|\tilde{Y}, M_2)\}^{1/T_i},$$
3. assign random initial values  $\psi_{i,0} = (p_{i0}, r_{i0}, u_{i0})$  to each MCMC chain; set  $t = 0$ ,
4. update  $\psi_{i,t}$  to  $\psi_{i,t+1}$  by Gibbs sampler as explained in subsection 2.3.1.2, and exchange  $\psi_{i,t+1}$  with its neighbor(s) as explained in subsection 2.3.1.1, for  $i = 1, \dots, N$ ; set  $t = t + 1$ ,
5. repeat the last step until  $t = K$ ; check the convergence and mixing of  $\{\psi_{N,t}\}_{t=0}^K$ ; adjust the temperature ladder and repeat the above steps until the MCMC chain converges and mixes well,
6. discard the first 20% of the sequence  $\{\psi_{N,t}\}_{t=0}^K$  (called burn-in), take every 10 samples from the rest of the chain (called thinning), denote the new sequence by  $\{\psi_j\}$ , which are samples from the posterior distribution,
7. calculate the Bayes factor estimator  $\widehat{\text{BF}}$  as in equation (2.8), and the posterior sample mean  $\widehat{\psi} = (\widehat{p}, \widehat{r}, \widehat{u})$ ,
8. conclude the presence of the source if  $\widehat{\text{BF}} > 3$ ; the source strength (i.e. the emission rate) is estimated by  $\widehat{p}$ ; if  $\widehat{\text{BF}} > 3$ , the location of the detected source is estimated by  $\{\widehat{r} \cos(\widehat{u}), \widehat{r} \sin(\widehat{u})\}$ .

Furthermore, the uncertainty in estimation could be summarized by other statistics such as sample standard deviation.

## 2.4. Simulation Study

We considered the situation where the size of the possible source is approximately known and is small compared to the size of the whole object. After choosing appro-

priate units, we assume that the object is the unit disk. The practically reasonable assumption is that the source radius is around 1% of the object radius, i.e.  $d = 0.01$  (e.g., the object has dimension of several meters, while the source is of diameter of a few centimeters). The simulation is designed to examine the performance of the method at various emission rate levels, which are chosen as  $p = 0.01$ ,  $p = 0.005$ ,  $p = 0.001$  and the case that no source exists,  $p = 0.00$ . We experimented with the number of detected particles being  $n = 2 \times 10^5$  and  $n = 5 \times 10^5$ . With a fixed  $d$ , as the true emission rate  $p$  (and thus signal-to-noise ratio) decreases, a larger total number of all detected particles is required for detection of a source. This can be explained by a simple application of the Central Limit Theorem CLT (see, for example, Allmaras et al., 2010). The prior values  $[0 < p_1, \dots, p_h]$  are assumed to be located near the true value of  $p$ , which in many applications is known with some uncertainty. At each level  $p$ , 10 simulated data sets were generated and analyzed, including also the case  $p = 0$ . The results with two different sample sizes are summarized in Table 1 and Table 2, respectively. Along with the Bayes factors, we also report the posterior probability that there is no sources, namely

$$\begin{aligned} \text{pr}(p = 0 | \tilde{Y}) &= \text{pr}(M_1 | \tilde{Y}) \\ &= \frac{\text{pr}(M_1) \text{pr}(\tilde{Y} | M_1)}{\text{pr}(M_1) \text{pr}(\tilde{Y} | M_1) + \text{pr}(M_2) \text{pr}(\tilde{Y} | M_2)} = (1 + \text{BF})^{-1}, \end{aligned}$$

where BF refers to the Bayes factor.

One can note from the results that if  $p = 0.005$  or  $p = 0.01$ , much smaller sample sizes are sufficient to detect the source. As the level decreases, say for  $p = 0.0005$ , much larger sample sizes are required. In particular, the rows with  $p = 0.001$  of the Table 1 show that sensitivity is not too high. The reason is that the number of detected particles,  $n = 2 \times 10^5$ , is not high enough. The next table shows significant

improvements when the number of particles is increased.

Table 1. Summary of Bayes factors for simulation in Section 2.4. Sample size  $n = 2 \times 10^5$ . There are 10 simulations performed at each combination of level  $p$  and location, and 20 simulations at  $p = 0$ . The values reported in the table are minimum, median, maximum of 10 Bayes factors, and the proportion of Bayes factors being greater than 3. In the last column is the median of  $\text{pr}(p = 0|\tilde{Y})$  calculated from the 10 data sets.

$p$	Location	Min	Med	Max	Prop>3	$\text{pr}(p = 0 \tilde{Y})$
0.01	(0.6, 0.3)	$3.7 \times 10^{282}$	Inf	Inf	1	0
0.01	(0.96, -0.1)	$6.7 \times 10^{297}$	Inf	Inf	1	0
0.005	(0.6, 0.3)	$2.8 \times 10^{75}$	$3.1 \times 10^{89}$	$2.4 \times 10^{99}$	1	$3.1 \times 10^{-90}$
0.005	(0.96, -0.1)	$3.5 \times 10^{76}$	$1.2 \times 10^{86}$	$4.6 \times 10^{106}$	1	$8.1 \times 10^{-87}$
0.001	(0.6, 0.3)	0.3299	11.2	$7.7 \times 10^5$	0.6	$8.1 \times 10^{-2}$
0.001	(0.96, -0.1)	0.1032	1.4	$2.1 \times 10^5$	0.4	0.4051
0	N/A	0.36	1.06	24.73	0.1	0.48

Table 2. Repeat of the Table 1 with  $5 \times 10^5$  samples.

$p$	Location	Min	Med	Max	Prop>3	$\text{pr}(p = 0 \tilde{Y})$
0.01	(0.6,0.3)	Inf	Inf	Inf	1	0
0.01	(0.96, -0.1)	Inf	Inf	Inf	1	0
0.005	(0.6,0.3)	$1.02 \times 10^{219}$	$9.81 \times 10^{235}$	$7.77 \times 10^{275}$	1	$6.48 \times 10^{-236}$
0.005	(0.96, -0.1)	$2.00 \times 10^{214}$	$4.13 \times 10^{223}$	$1.91 \times 10^{255}$	1	$3.35 \times 10^{-223}$
0.001	(0.6,0.3)	523.76	$1.41 \times 10^9$	$5.44 \times 10^{16}$	1	$7.32 \times 10^{-10}$
0.001	(0.96, -0.1)	41.68	$2.35 \times 10^9$	$1.60 \times 10^{12}$	1	$4.96 \times 10^{-10}$
0	n/a	0.65	1.16	2.23	0	0.46

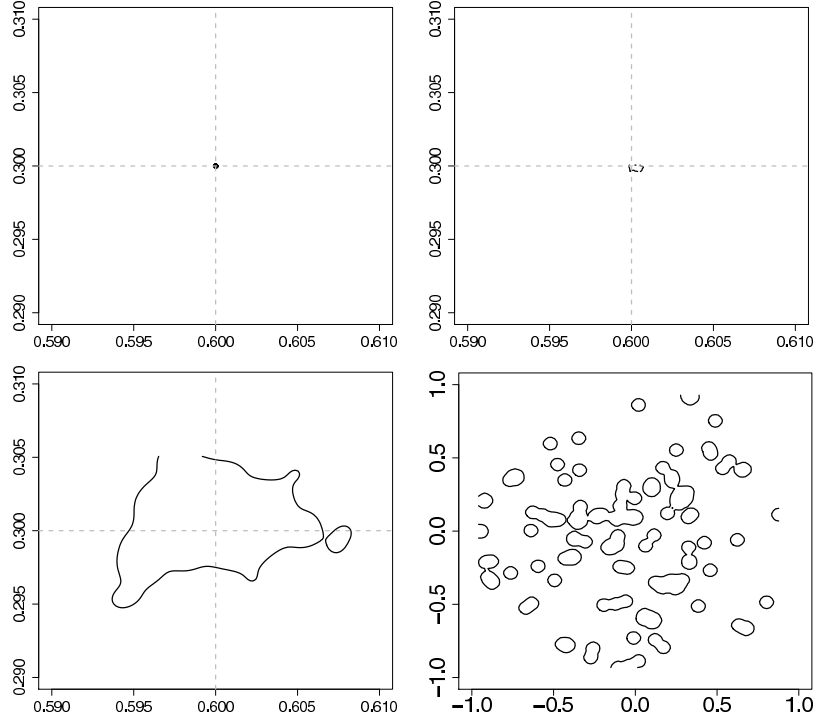


Figure 3. The estimated location of the source with  $n = 5 \times 10^5$  sample counts for various emitting levels, with 95% highest posterior density region. The top left plot is with  $p = 0.01$ ; the top right plot is with  $p = 0.005$ ; the bottom left plot is with  $p = 0.001$ ; the bottom right plot is with  $p = 0$ . In the plots where a source exists, the location of its center is indicated by the intersection of gray dashed lines. Each of the above figures is plotted with the posterior sample having median Bayes factors among the 10 simulated cases.

The results in Table 2 clearly show very high sensitivity to the presence of a source, which is indicated by the overall large values of BF. Furthermore, the location, if the source is present, can be also found with high accuracy, as shown in Figure 3.

## 2.5. Concluding Remarks

- The results of this study show that Bayesian methods can be successfully used for detection of low emission small sources in the cases of realistic parameters. High sensitivity can be achieved if the observation time (and thus total count of particles detected) is sufficiently large.
- The assumption that the detector can determine the directional information is crucial. We believe that otherwise detection with such low values of SNR (signal-to-noise ratio) would be impossible. The determination of the incoming direction is usually achieved by detector collimation. This is not an option with the extremely low SNR levels, since it would most probably eliminate completely the useful signal. However, there exist the so called Compton type cameras for detecting  $\gamma$ -photons (e.g., Allmaras et al., 2010, LeBlanc et al., 1998 and references therein), as well as their analogs (although based upon a somewhat different physics) for neutron detection (Marianno et al., 2010; Spence and Charlton, 2009). These cameras do not use collimation, but can determine some less precise directional information. Namely, the camera is able to provide a (hollow) cone of possible directions of the incoming particle. Although this is a less precise (and highly over-determined) information, it is known (e.g., see Allmaras et al., 2010 and references therein) how to convert the Compton type data into the precise directional information. Thus, the algorithms described can be used in conjunction with Compton type detectors.
- The goal of this article was to test the suggested detection technique in principle. This is why many simplifying and not exceedingly realistic assumptions were made (e.g., uniformity of the background, no scattering effects, precise de-



termination of the direction of an incoming particle, etc.). The assumption that the scattered emission from the source is considered as a part of the background and thus is not addressed directly, should not matter much, due to the small size of this emission. This is confirmed by some numerical experiments. Other simplifying assumptions might not be that benign, though.

- Now, when the workability of the algorithm is shown in the simplest situation, it is planned to address in the future study the effects of scattering, inhomogeneous random background noise, Compton type cameras, as well as the  $3D$  situation. It is also planned to compare the results and the computational cost of the Bayesian approach with the more analytic techniques of Allmaras et al. (2010).

## CHAPTER III

PARAMETER ESTIMATION OF PARTIAL DIFFERENTIAL EQUATION  
MODELS**3.1. Introduction**

Differential equations are important tools in modeling dynamic processes, and are widely used in many areas. The forward problem of solving equations or simulating state variables for given parameters that define the differential equation models has been studied extensively by mathematicians. However, the inverse problem of estimating parameters based on observed error-prone state variables has a relatively sparse statistical literature, and this is especially the case for partial differential equation (PDE) models. There is growing interest in developing efficient estimation methods for such problems.

Various statistical methods have been developed to estimate parameters in ordinary differential equation (ODE) models. There is a series of work in the study of HIV dynamics in order to understand the pathogenesis of HIV infection. For example, Ho et al. (1995) and Wei et al. (1995) used standard nonlinear least square regression methods; Wu, Ding and DeGruttola (1998) and Wu and Ding (1999) first proposed a mixed-effects model approach. Refer to Wu (2005) for a comprehensive review of these methods. Furthermore, Putter et al. (2002), Huang and Wu (2006), and Huang, Liu and Wu (2006) proposed hierarchical Bayesian approaches for this problem. These methods require repeatedly solving ODE models numerically, which could be time-consuming. Ramsay (1996) proposed a data reduction technique in functional data analysis which involved solving for coefficients of linear differential operators, see Poyton et al. (2006) for an example of application. Li et al. (2002)

studied a pharmacokinetic model and proposed a semiparametric approach for estimating time-varying coefficients in an ODE model. Ramsay et al. (2007) proposed a generalized smoothing approach, named parameter cascading, for estimating constant parameters in ODE models, based on data smoothing methods and profiled estimation. Cao, Wang and Xu (2011) proposed robust estimation for ODE models when data have outliers. Cao, Huang and Wu (2011) proposed a parameter cascading method to estimate time-varying parameters in ODE models. These methods estimate parameters by optimizing certain criteria. In the optimization procedure, using gradient-based optimization techniques may have the parameter estimates converge to a local minima, otherwise global optimization is computationally intensive.

Another strategy to estimate parameters of ODE is the two-stage method, which in the first stage estimate the function and its derivatives from noisy observations using data smoothing methods without considering differential equation models, and then in the second stage estimates of ODE parameters are obtained following the least squares principle. Liang and Wu (2008) developed a two-stage method for a general first order ODE model, using local polynomial regression in the first stage, and established asymptotic properties of the proposed estimator under the framework of measurement error models. Similarly, Chen and Wu (2008) developed local estimation for time-varying coefficients. The two-stage methods are easy to implement, however, they might not be statistically efficient, due to the fact that derivatives cannot be estimated accurately from noisy data, especially higher order derivatives.

As for PDE, there are two main approaches. The first is similar to the two-stage method in Liang and Wu (2008). For example, Bar, Hegger and Kantz (1999) modeled unknown PDEs using multivariate polynomials of sufficiently high order, and the best fit was chosen by minimizing the least squares error of the polynomial approximation. Based on estimated functions, the PDE parameters were estimated

using the least square principle (Muller and Timmer, 2004). The issues of noise level and data resolution were extensively addressed in this approach. See also Parlitz and Merkwirth (2000) and Voss et al. (1999) for more examples. The second approach uses numerical solutions of PDEs, thus circumventing derivative estimation. For example, Muller and Timmer (2002) solved the target least squares-type minimization problem using an extended multiple shooting method. The main idea was to solve initial value problems in sub-intervals and integrate the segments with additional continuity constraints. Global minima can be reached in this algorithm, but it requires careful parameterization of the initial condition, and the computational cost is high.

In this article, we consider a multidimensional dynamic process,  $g(\mathbf{x})$ , where  $\mathbf{x} = (x_1, \dots, x_p)^T \in \mathbb{R}^p$  is a multi-dimensional argument. Suppose this dynamic process can be modeled with a PDE model

$$\mathcal{F}\left(\mathbf{x}, g, \frac{\partial g}{\partial x_1}, \dots, \frac{\partial g}{\partial x_p}, \frac{\partial^2 g}{\partial x_1 \partial x_1}, \dots, \frac{\partial^2 g}{\partial x_1 \partial x_p}, \dots; \boldsymbol{\theta}\right) = 0, \quad (3.1)$$

where  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)^T$  is the parameter vector of primary interest, and the left hand side of (3.1) has a parametric form in  $g(\mathbf{x})$  and its partial derivatives. In practice, we might not observe  $g(\mathbf{x})$  but its surrogate  $Y(\mathbf{x})$ . We assume that  $g(\mathbf{x})$  is observed over a meshgrid with measurement errors, so that for  $i = 1, \dots, n$ , we observe data  $(Y_i, \mathbf{x}_i)$  satisfying

$$Y_i = g(\mathbf{x}_i) + \epsilon(\mathbf{x}_i),$$

where  $E\{\epsilon(\mathbf{x})\} = 0$  and  $\text{var}\{\epsilon(\mathbf{x})\} = \sigma_\epsilon^2(\mathbf{x})$ . Our goal is to estimate the unknown  $\boldsymbol{\theta}$  in PDE model (3.1) from noisy data, and to quantify the uncertainty of the estimates.

As mentioned before, a straightforward two-stage strategy, though easy to implement, has difficulty in estimating derivatives accurately, and the ODE parameter estimates are biased. We propose two joint modeling schemes: (a) a parameter cascad-

ing approach and (b) a fully Bayesian treatment. We conjecture that joint modeling approaches are more statistically efficient than a two-stage method, a conjecture that is borne out in our simulations.

The main idea of our two methods is to represent the unknown dynamic process via a nonparametric function while using the PDE model to regularize the fit. In both methods, the nonparametric function is expressed as a linear combination of B-spline basis functions. In the parameter cascading method, this nonparametric function is estimated using the penalized least squares principle, where a penalty term is defined to incorporate the PDE model. This penalizes the infidelity of the nonparametric function to the PDE model, so that the nonparametric function is forced to better represent the dynamic process modeled by the PDE. In the Bayesian method, the PDE model information is coded in the prior distribution. We recognize that there is no exact solution by substituting the nonparametric function into the PDE model (3.1). This PDE modeling error is then modeled as a random process, hence inducing a constraint on the basis function coefficients. We also introduce in the prior an explicit penalty on the smoothness of the nonparametric function. Our two methods avoid direct estimation of the derivative of the dynamic process, but it can be obtained easily as a linear combination of the derivatives of the basis functions.

In principle, the proposed methods are applicable to all PDE, thus having potentially wide applications. As quick examples of PDE, the heat equation and wave equation are among the most famous ones. The heat equation, also known as the diffusion equation, describes the evolution in time of the heat distribution or chemical concentration in a given region, and is defined as  $\partial g(\mathbf{x}, t)/\partial t - \theta \sum_{i=1}^p \partial^2 g(\mathbf{x}, t)/\partial x_i^2 = 0$ . The wave equation is a simplified model for description of waves, such as sound waves, light waves and water waves, and is defined as  $\partial^2 g(\mathbf{x}, t)/\partial t^2 = \theta^2 \sum_{i=1}^p \partial^2 g(\mathbf{x}, t)/\partial x_i^2$ . More examples of famous PDE are the Laplace equation, the transport equation and

the beam equation. Please refer to Evans (1998) for a detailed introduction of PDE.

For illustration, we will do specific calculations based on our empirical example of LIDAR data described in Section 3.6 and also used in our simulations in Section 3.5. There we propose a PDE model for received signal  $g(t, z)$  over time  $t$  and range  $z$  given as

$$\partial g(t, z)/\partial t - \theta_D \partial^2 g(t, z)/\partial z^2 - \theta_S \partial g(t, z)/\partial z - \theta_A g(t, z) = 0. \quad (3.2)$$

The above PDE has a closed form solution, obtained by separation of variables, but the solution is the sum of an infinite sequence. It requires a high computational load to evaluate the solution over a meshgrid of moderate size.

The rest of the chapter is organized as follows. The basic idea of basis function approximation is explained in Section 3.2. The parameter cascading method is introduced in Section 3.3, and the asymptotic properties of the proposed estimator are established. In Section 3.4 we introduce the Bayesian framework and explain how to make posterior inference using the MCMC technique. Simulation studies are presented in Section 3.5 to evaluate the finite sample performance of our two methods in comparison with a two-stage method. In Section 3.6 we illustrate the methods using a LIDAR data from threat detection experiment. Finally, we conclude with some remarks in Section 3.7.

### 3.2. Basis Function Approximation

When solving partial differential equations, we are able to obtain a unique, explicit formula for certain specific examples, such as the wave equation. However, most PDEs used in practice have no explicit solutions. Then the PDE can only be solved with some numeric methods, such as finite difference method (Morton and Mayers,

2005) and finite element method (Brenner and Scott, 2010). Instead of repeatedly solving PDE numerically for thousands of parameter candidates, which is computationally expensive, we represent the dynamic process,  $g(\mathbf{x})$ , modeled in (3.1), by a nonparametric function, which can be expressed as a linear combination of basis functions

$$g(\mathbf{x}) = \sum_{k=1}^K b_k(\mathbf{x})\beta_k = \mathbf{b}^T(\mathbf{x})\boldsymbol{\beta}, \quad (3.3)$$

where  $\mathbf{b}(\mathbf{x}) = (b_1(\mathbf{x}), \dots, b_K(\mathbf{x}))^T$  is the vector of basis functions, and  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_K)^T$  is the vector of corresponding coefficients.

With approximation (3.3), the PDE model (3.1) can be represented using the same set of basis functions by substituting approximation (3.3) into model (3.1), so that

$$\mathcal{F}[\mathbf{x}, \mathbf{b}^T(\mathbf{x})\boldsymbol{\beta}, \{\partial\mathbf{b}(\mathbf{x})/\partial x_1\}^T\boldsymbol{\beta}, \dots; \boldsymbol{\theta}]. \quad (3.4)$$

In the special case of linear PDEs, the above expression is also linear in  $\boldsymbol{\beta}$ . We write the approximated linear PDE model as

$$\mathcal{F}[\mathbf{x}, \mathbf{b}^T(\mathbf{x})\boldsymbol{\beta}, \{\partial\mathbf{b}(\mathbf{x})/\partial x_1\}^T\boldsymbol{\beta}, \dots; \boldsymbol{\theta}] = \mathbf{f}^T\{\mathbf{b}(\mathbf{x}), \partial\mathbf{b}(\mathbf{x})/\partial x_1, \dots; \boldsymbol{\theta}\}\boldsymbol{\beta}, \quad (3.5)$$

where  $\mathbf{f}\{\mathbf{b}(\mathbf{x}), \partial\mathbf{b}(\mathbf{x})/\partial x_1, \dots; \boldsymbol{\theta}\}$  is a linear function of the basis functions and their derivatives. In the following context, we denote  $\mathbf{f}\{\mathbf{b}(\mathbf{x}), \partial\mathbf{b}(\mathbf{x})/\partial x_1, \dots; \boldsymbol{\theta}\}$  by the short hand notation  $\mathbf{f}(\mathbf{x}; \boldsymbol{\theta})$ . For the PDE example (3.2), the form of  $\mathbf{f}(\mathbf{x}; \boldsymbol{\theta})$  is given in Appendix A.1.

It seems that model (3.3) represents the infinite-dimensional space of functions using a framework of fixed dimension  $K$ . However, it is a great deal how the basis functions are chosen. When  $K$  is equal to the number of observations, interpolation is achieved exactly, and as the number of basis is reduced, extra smoothness is intro-

duced. Hence, the number of basis  $K$  itself is a parameter that we need to decide based on characteristics of the data. It is possible to choose knots automatically. One might naturally think about model selection criterion, such as cross-validation. But the number of model fits increases rapidly as the number of candidate knots increases. Then this approach easily becomes impossible. Many of the feasible frequentist methods, for example, Friedman and Silverman (1989) and Stone et al. (1997), are based on stepwise regression. Bayesian framework is also available, see Denison, Mallick and Smith (1997) for example. Wand (2000) provided a comprehensive review of some knots selection approaches. Despite of good performance, knots selection procedures are highly computationally intensive.

Typically, the basis functions are nonlinear transformations of the data, hence model (3.3) has the flexibility to model data adequately. In principle, change of basis does not change the model fit result. Ideally, we could use basis functions which have matching features to the unknown functions. Usually in this way, a relatively small number of basis functions are required to achieve an approximation of similar degree of satisfaction. For example, the Fourier basis system would be a natural choice for periodic data. However, there are other considerations including numerical stability, computational cost, ease and accuracy of implementation and interpretability, etc. For example, it is certainly undesirable to invert a nearly rank-deficient matrix during the computation.

The choice of basis is especially important when one wants to estimate derivatives of the fit. It happens that basis working well for function estimation results in poor derivative estimation. Besides Fourier basis, other commonly used basis include truncated power basis, B-spline basis, and radial basis functions. Eilers and Marx (2010) discussed the close relationship between truncated power basis and B-spline basis. While B-spline basis can be constructed from truncated power basis by com-



puting repeated differences, the B-splines are much more orthogonal, leading to more computational efficiency and numerical stability. And the advantage of radial basis functions is that the extension to higher dimension is straightforward.

### 3.2.1. B-Spline Basics

We choose B-splines as basis functions in all simulations and applications in this work, since B-splines are non-zeros only in short subintervals, a feature called the compact support property (de Boor, 2001), which is very useful for efficient computation and numerical stability, compared with other basis (e.g. truncated power basis). The B-spline basis functions are defined with their order, the number of knots and their locations (one such building block is shown in Figure 4). To avoid the complicated knot selection problem, we use a large enough number of knots to make sure the basis functions are sufficiently flexible to approximate the dynamic process. A rule of thumb is to put one knot at each data point, so users do not have to select the number of knots and their locations. To prevent the nonparametric function overfitting the data, one penalty term will be defined with the PDE model in the next section to penalize the roughness of the nonparametric function.

### 3.3. Parameter Cascading Method

The parameter cascading method is a generalized smoothing approach and multi-criteria optimization procedure. There are three nested levels of optimization with respect to the basis coefficient, the PDE parameter and a smoothing parameter. This nested structure leads to the notation of parameter cascading (Ramsay et al., 2007). The algorithm is introduced in the follow subsections.

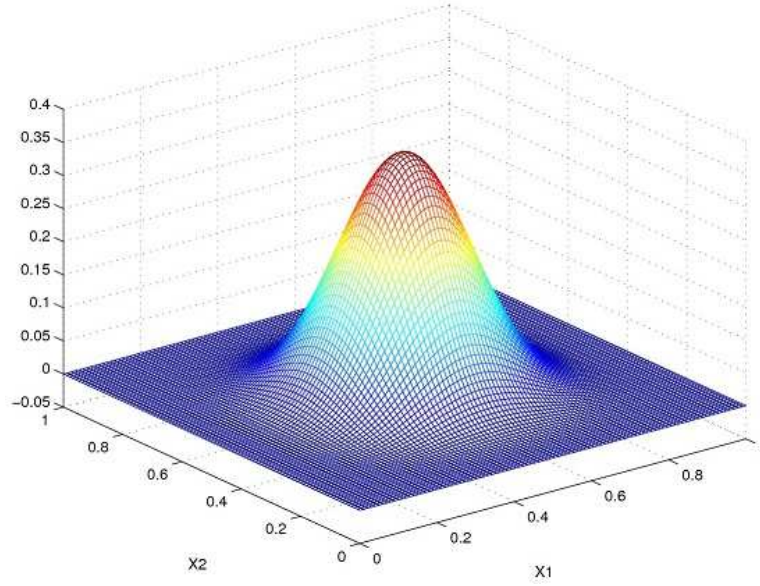


Figure 4. Example of building block of B-splines. This is a 2D quartic B-spline basis function formed by tensor product of 1D quartic B-spline functions, at knot  $(0.5, 0.5)$ .

### 3.3.1. Estimating PDE Parameters

Following Section 3.2, the dynamic process  $g(\mathbf{x})$  is expressed as a linear combination of basis functions. It is natural to estimate the basis function coefficients  $\boldsymbol{\beta}$  using penalized splines (Eilers and Marx, 2010; Ruppert, Wand and Carroll, 2003). If we were simply interested in estimating  $g(\cdot) = \mathbf{b}^T(\cdot)\boldsymbol{\beta}$ , then we would use the usual penalty  $\lambda\boldsymbol{\beta}^T\mathbf{P}^T\mathbf{P}\boldsymbol{\beta}$ , where  $\lambda$  is a penalty parameter and  $\mathbf{P}$  is a matrix performing difference on adjacent elements of  $\boldsymbol{\beta}$  (Eilers and Marx, 2010). Such a penalty does penalize on the smoothness of the estimated function, however, it is not in fidelity with (3.1). Instead, for fixed  $\boldsymbol{\theta}$ , we define the roughness penalty as  $\int[\mathcal{F}\{\mathbf{x}, g(\mathbf{x}), \dots; \boldsymbol{\theta}\}]^2 d\mathbf{x}$ . This penalty incorporates the PDE model, containing derivatives involved in the model. As a result, the penalty is able to regularize the spline fit. It also shows fidelity to the PDE model, i.e., smaller value indicates more fidelity of the spline

approximation to the PDE. Hence, we propose to estimate the coefficients  $\beta$  for fixed  $\theta$  by minimizing the penalized least squares

$$J(\beta|\theta) = \sum_{i=1}^n \{Y_i - g(\mathbf{x}_i)\}^2 + \lambda \int [\mathcal{F}\{\mathbf{x}, g(\mathbf{x}), \dots; \theta\}]^2 d\mathbf{x}. \quad (3.6)$$

The integration in (3.6) can be approximated numerically by quadrature. Burden and Douglas (2010) suggested that the composite Simpson's rule provide an adequate approximation. See Appendix A.2 for details.

The PDE parameter  $\theta$  is then estimated at an upper level of optimization. Denote the estimate of the spline coefficients by  $\hat{\beta}(\theta)$ , which is considered as a function of  $\theta$ . Define  $\hat{g}(\mathbf{x}, \theta) = \mathbf{b}^T(\mathbf{x})\hat{\beta}(\theta)$ . As the estimator  $\hat{\beta}(\theta)$  is already regularized, we propose to estimate  $\theta$  by minimizing the following straightforward measure of fit

$$H(\theta) = \sum_{i=1}^n \{Y_i - \hat{g}(\mathbf{x}_i, \theta)\}^2 = \sum_{i=1}^n \{Y_i - \mathbf{b}^T(\mathbf{x}_i)\hat{\beta}(\theta)\}^2. \quad (3.7)$$

For a general nonlinear PDE model, the function  $\hat{\beta}(\theta)$  might not have a close form, and the estimate is thus obtained numerically. This lower level of optimization for fixed  $\theta$  is embedded inside the optimization of  $\theta$ . The objective functions  $J(\beta|\theta)$  and  $H(\theta)$  are minimized iteratively until convergence of the solution. In some cases, the optimization could be accelerated and made more stable by providing the gradient, whose analytic form, by the chain rule, is

$$\frac{\partial H(\theta)}{\partial \theta} = \left\{ \frac{\partial \hat{\beta}(\theta)}{\partial \theta} \right\}^T \frac{\partial H(\theta)}{\partial \hat{\beta}(\theta)}.$$

Although  $\hat{\beta}(\theta)$  has no explicit expression, the implicit function theorem can be applied to find the analytic form of the first-order derivative of  $\hat{\beta}(\theta)$  with respect to  $\theta$  required in the above gradient. As  $\hat{\beta}$  is the minimizer of  $J(\beta|\theta)$ , we have  $\partial J(\beta|\theta)/\partial \beta|_{\hat{\beta}} = 0$ . By taking the total derivative with respect to  $\theta$  on the left hand side, and assuming

$\partial^2 J(\boldsymbol{\beta}|\boldsymbol{\theta})/\partial\boldsymbol{\beta}^T\partial\boldsymbol{\beta}|_{\hat{\boldsymbol{\beta}}}$  is non-singular, the analytic expression of the first-order derivative of  $\hat{\boldsymbol{\beta}}$  is

$$\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\boldsymbol{\theta}} = - \left( \frac{\partial^2 J}{\partial\boldsymbol{\beta}^T\partial\boldsymbol{\beta}} \Big|_{\hat{\boldsymbol{\beta}}} \right)^{-1} \left( \frac{\partial^2 J}{\partial\boldsymbol{\theta}^T\partial\boldsymbol{\beta}} \Big|_{\hat{\boldsymbol{\beta}}} \right). \quad (3.8)$$

When the PDE model (3.1) is linear,  $\hat{\boldsymbol{\beta}}$  has a close form and the algorithm could be stated as follows. By substituting in (3.3) and (3.5), the lower level criterion (3.6) becomes

$$J(\boldsymbol{\beta}|\boldsymbol{\theta}) = \sum_{i=1}^n \{Y_i - \mathbf{b}^T(\mathbf{x}_i)\boldsymbol{\beta}\}^2 + \lambda \int \boldsymbol{\beta}^T \mathbf{f}(\mathbf{x}; \boldsymbol{\theta}) \mathbf{f}^T(\mathbf{x}; \boldsymbol{\theta}) \boldsymbol{\beta} d\mathbf{x}.$$

Let  $\mathbf{B}$  be the  $n \times K$  basis matrix with  $i^{th}$  row  $\mathbf{b}^T(\mathbf{x}_i)$ , and define  $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ , and the  $K \times K$  penalty matrix  $\mathbf{R}(\boldsymbol{\theta}) = \int \mathbf{f}(\mathbf{x}; \boldsymbol{\theta}) \mathbf{f}^T(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x}$ . See Appendix A.2 for calculation of  $\mathbf{R}(\boldsymbol{\theta})$  for the PDE example (3.2). Then the penalized least square criterion (3.6) can be expressed in the matrix notation

$$J(\boldsymbol{\beta}|\boldsymbol{\theta}) = (\mathbf{Y} - \mathbf{B}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{B}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^T\mathbf{R}(\boldsymbol{\theta})\boldsymbol{\beta}, \quad (3.9)$$

which is a quadratic function of  $\boldsymbol{\beta}$ . By minimizing the above penalized least square criterion, the estimate for  $\boldsymbol{\beta}$ , for fixed  $\boldsymbol{\theta}$ , can be obtained in a close formula as follows

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \{\mathbf{B}^T\mathbf{B} + \lambda\mathbf{R}(\boldsymbol{\theta})\}^{-1}\mathbf{B}^T\mathbf{Y}.$$

Then by substituting in the above estimator, the upper level criterion (3.7) becomes

$$H(\boldsymbol{\theta}|\lambda) = \|\mathbf{Y} - \mathbf{B}\{\mathbf{B}^T\mathbf{B} + \lambda\mathbf{R}(\boldsymbol{\theta})\}^{-1}\mathbf{B}^T\mathbf{Y}\|^2. \quad (3.10)$$

To summarize, when estimating parameters in linear PDE models, we minimize criterion (3.10) to obtain an estimate,  $\hat{\boldsymbol{\theta}}$ , for parameters in linear PDE models. The estimated basis coefficients,  $\hat{\boldsymbol{\beta}}$ , is obtained by substituting  $\hat{\boldsymbol{\theta}}$  into (3.10).

### 3.3.2. Smoothing Parameter Selection

Our ultimate goal is to obtain an estimate for the PDE parameter  $\boldsymbol{\theta}$  such that the solution of the PDE is close to the observed data. For any given value of the smoothing parameter,  $\lambda$ , we obtain the PDE parameter estimate,  $\widehat{\boldsymbol{\theta}}$ , and the basis coefficient estimate,  $\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\theta}})$ . Hence, both of them can be treated as functions of  $\lambda$ , which are denoted as  $\widehat{\boldsymbol{\theta}}(\lambda)$  and  $\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\theta}}(\lambda), \lambda)$ . Let define  $e_i(\lambda) = Y_i - \widehat{g}\{\mathbf{x}_i, \widehat{\boldsymbol{\theta}}(\lambda), \lambda\}$  and  $\eta_i(\lambda) = \mathcal{F}[\widehat{g}\{\mathbf{x}_i, \widehat{\boldsymbol{\theta}}(\lambda)\}, \widehat{\boldsymbol{\theta}}(\lambda)]$ , the latter of which is  $\widehat{\mathbf{f}}^T\{\mathbf{x}_i; \widehat{\boldsymbol{\theta}}(\lambda)\}\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\theta}}(\lambda), \lambda)$  for linear PDE models. Fidelity to the PDE can be measured by  $\sum_{i=1}^n \eta_i^2(\lambda)$ , while fidelity to the data can be measured by  $\sum_{i=1}^n e_i^2(\lambda)$ . Clearly, minimizing just  $\sum_{i=1}^n e_i^2(\lambda)$  leads to  $\lambda = 0$ , and gives far too undersmoothed data fits, while not taking the PDE into account. On the other hand, our experience shows that minimizing simply  $\sum_{i=1}^n \eta_i^2(\lambda)$  always results in the largest candidate value for  $\lambda$ .

Hence, we propose the following criterion, which considers data fitting and PDE model fitting simultaneously. To choose an optimal  $\lambda$ , we minimize

$$G(\lambda) = \sum_{i=1}^n e_i^2(\lambda) + \sum_{i=1}^n \eta_i^2(\lambda). \quad (3.11)$$

### 3.3.3. Variance Estimation of Parameters

In this section, we derive and justify a sandwich estimator of the covariance matrix of the PDE parameter estimator  $\widehat{\boldsymbol{\theta}}$  for fixed  $\lambda$ . For notational convenience we thus drop the dependence of  $\boldsymbol{\theta}$  and  $\boldsymbol{\beta}$  on  $\lambda$ . The parameter cascading method estimates the PDE parameter,  $\boldsymbol{\theta}$ , by solving  $\partial H(\boldsymbol{\theta})/\partial \boldsymbol{\theta} = 0$ , which is the estimating equation

$$\frac{\partial H(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -2 \sum_{i=1}^n \{Y_i - \mathbf{b}^T(\mathbf{x}_i)\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})\} \left( \frac{\partial \widehat{\boldsymbol{\beta}}}{\partial \boldsymbol{\theta}^T} \right)^T \mathbf{b}(\mathbf{x}_i) = -2 \sum_{i=1}^n \Psi_i(\boldsymbol{\theta}) = 0.$$

where  $\Psi_i = \{Y_i - \mathbf{b}^T(\mathbf{x}_i)\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})\}(\partial \widehat{\boldsymbol{\beta}}/\partial \boldsymbol{\theta}^T)^T \mathbf{b}(\mathbf{x}_i)$ . Qi and Zhao (2010) shows that the parameter cascading estimate,  $\widehat{\boldsymbol{\theta}}$ , is a consistent estimator of  $\boldsymbol{\theta}$ . Let  $\boldsymbol{\theta}_0$  denote the

true value of the PDE parameter. Define the score for  $\boldsymbol{\theta}$  as  $\mathcal{S}_n(\boldsymbol{\theta}) = n^{-1/2} \sum_{i=1}^n \Psi_i(\boldsymbol{\theta})$ . Doing a Taylor series and assuming  $n^{-1/2}$ -convergence of  $\widehat{\boldsymbol{\theta}}$ , we have that  $\mathbf{0} = \mathcal{S}_n(\widehat{\boldsymbol{\theta}}) = \mathcal{S}_n(\boldsymbol{\theta}_0) + n^{-1/2} \mathcal{M}_n(\boldsymbol{\theta}_0) n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_p(1)$ , where  $\mathcal{M}_n(\boldsymbol{\theta}) = \partial \mathcal{S}_n(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}^T$  is the Hessian matrix. Applying the law of large numbers, we have  $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \approx -\boldsymbol{\Lambda}_n^{-1}(\boldsymbol{\theta}_0) \sum_{i=1}^n \Psi_i(\boldsymbol{\theta}_0)$ , where  $\boldsymbol{\Lambda}_n(\boldsymbol{\theta}_0) = \sum_{i=1}^n E\{\partial \Psi_i(\boldsymbol{\theta}_0) / \partial \boldsymbol{\theta}^T\}$ .

Define  $\tilde{\lambda} = \lambda/n$ , and matrices

$$\begin{aligned} \mathbf{S}_n &= n^{-1} \sum_{i=1}^n \mathbf{b}(\mathbf{x}_i) \mathbf{b}^T(\mathbf{x}_i), \\ \boldsymbol{\Lambda}_n(\widehat{\boldsymbol{\theta}}) &= \sum_{i=1}^n \partial \Psi_i(\widehat{\boldsymbol{\theta}}) / \partial \boldsymbol{\theta}^T, \\ \mathbf{G}_n(\boldsymbol{\theta}) &= \mathbf{S}_n + \tilde{\lambda} \mathbf{R}(\boldsymbol{\theta}), \\ \mathbf{R}_{j\theta}(\boldsymbol{\theta}) &= \partial \mathbf{R}(\boldsymbol{\theta}) / \partial \theta_j, \\ \widehat{\mathcal{V}}_j &= \mathbf{R}(\widehat{\boldsymbol{\theta}}) \mathbf{G}_n^{-1}(\widehat{\boldsymbol{\theta}}) \mathbf{R}_{j\theta}(\widehat{\boldsymbol{\theta}}), \\ \widehat{\mathcal{W}}_j &= \widehat{\mathcal{V}}_j + \widehat{\mathcal{V}}_j^T, \end{aligned}$$

and  $\widehat{\mathcal{C}}_{jk} = n \tilde{\lambda}^4 \widehat{\sigma}_\epsilon^2 \widehat{\boldsymbol{\beta}}^T(\widehat{\boldsymbol{\theta}}) \widehat{\mathcal{W}}_j \mathbf{G}_n^{-1}(\widehat{\boldsymbol{\theta}}) \mathbf{S}_n \mathbf{G}_n^{-1}(\widehat{\boldsymbol{\theta}}) \widehat{\mathcal{W}}_k \widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\theta}})$ . We show in Appendix A.3 that  $n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$  is asymptotically normally distributed with mean zero and covariance matrix consistently estimated by

$$\boldsymbol{\Lambda}_n^{-1}(\widehat{\boldsymbol{\theta}}) \mathcal{C}(\widehat{\boldsymbol{\theta}}) \{\boldsymbol{\Lambda}_n^{-1}(\widehat{\boldsymbol{\theta}})\}^T,$$

where  $\mathcal{C}(\widehat{\boldsymbol{\theta}})$  is a matrix whose  $(j, k)^{th}$  element is  $\widehat{\mathcal{C}}_{jk}$ . Here  $\widehat{\sigma}_\epsilon^2$  is the estimated variance of  $\epsilon(\mathbf{x}_i)$  and can be calculated by first fitting a standard spline regression and then forming the residual variance.

### 3.4. Bayesian Estimation and Inference

In this section we introduce a Bayesian approach for estimating parameters in PDE models. In this Bayesian approach, the dynamic process modeled by the PDE model is again represented by a linear combination of B-spline basis functions. The coefficients to the basis functions are regularized through the prior, which contains the PDE model information. Therefore, data fitting and PDE fitting are incorporated into a joint model.

We use the same notations as before. With the basis function approximation (3.3), the basis function model for data fitting is

$$Y_i = \mathbf{b}^T(\mathbf{x}_i)\boldsymbol{\beta} + \epsilon_i, \text{ for } i = 1, \dots, n, \quad (3.12)$$

where the  $\epsilon_i$  are independent and identically distributed measurement errors and are assumed to follow a Gaussian distribution with mean zero and variance  $\sigma_\epsilon^2$ . The basis functions are chosen with the same rule introduced in the previous section.

In the conventional Bayesian P-splines (described in Section 3.4.1), the penalty term penalizes on the smoothness of the estimated function, for example, by controlling the size of second-order derivatives. Rather than using a single optimal smoothing parameter as in frequentist methods, our Bayesian approach performs a model mixing with respect to this quantity. In other words, many different spline models would provide plausible representation of data, and the Bayesian approach treats such model uncertainty through the prior distribution of the smoothing parameter.

In our problem, we know further that the underlying function satisfies a given PDE model. Naturally, this information should be coded into the prior distribution to regularize the fit. As we recognize that there may be no basis function approximation that exactly satisfies the PDE model (3.1), for the purposes of Bayesian computation,

we will treat the approximation error as random, and the PDE modeling errors are

$$\mathcal{F}\{\mathbf{x}_i, \mathbf{b}^T(\mathbf{x}_i)\boldsymbol{\beta}, \dots; \boldsymbol{\theta}\} = \zeta(\mathbf{x}_i), \quad (3.13)$$

where the random modeling errors,  $\zeta(\mathbf{x}_i)$ , are random, and are assumed to be independent and identically distributed with a prior distribution  $\text{Normal}(0, \gamma_0^{-1})$ , where the precision  $\gamma_0$  should be large enough so that the approximation error in solving (3.1) with a basis function approximation is small. Similarly, instead of using a single optimal value for the precision parameter,  $\gamma_0$ , a prior distribution is assigned to  $\gamma_0$ . The modeling error distribution assumption (3.13) and a roughness penalty constraint altogether induce a prior distribution on the basis function coefficients,  $\boldsymbol{\beta}$ , as

$$\begin{aligned} [\boldsymbol{\beta} | \boldsymbol{\theta}, \gamma_0, \gamma_1, \gamma_2] &\propto (\gamma_0 \gamma_1 \gamma_2)^{K/2} \exp\{-\gamma_0 \boldsymbol{\zeta}^T(\boldsymbol{\beta}, \boldsymbol{\theta}) \boldsymbol{\zeta}(\boldsymbol{\beta}, \boldsymbol{\theta})/2 \\ &\quad - \boldsymbol{\beta}^T(\gamma_1 H_1 + \gamma_2 H_2 + \gamma_1 \gamma_2 H_3) \boldsymbol{\beta}/2\}, \end{aligned} \quad (3.14)$$

where, as before,  $K$  denotes the number of basis functions,  $\gamma_0$  is the precision parameter,  $\boldsymbol{\zeta}(\boldsymbol{\beta}, \boldsymbol{\theta}) = [\mathcal{F}\{\mathbf{x}_1, \mathbf{b}^T(\mathbf{x}_1)\boldsymbol{\beta}, \dots; \boldsymbol{\theta}\}, \dots, \mathcal{F}\{\mathbf{x}_n, \mathbf{b}^T(\mathbf{x}_n)\boldsymbol{\beta}, \dots; \boldsymbol{\theta}\}]^T$ ,  $\gamma_1$  and  $\gamma_2$  control the amount of penalty on smoothness, and the penalty matrices  $H_1, H_2, H_3$  are the same as in the usual Bayesian P-spline, given in (3.16). We assume conjugate priors for  $\sigma_\epsilon^2$  and  $\gamma_\ell$  as  $\sigma_\epsilon^2 \sim \text{IG}(a_\epsilon, b_\epsilon)$ ,  $\gamma_\ell \sim \text{Gamma}(a_\ell, b_\ell)$ , for  $\ell = 0, 1, 2$ , where  $\text{IG}(a, b)$  denotes the Inverse-Gamma distribution with mean  $(a - 1)^{-1}b$ . For the PDE parameter,  $\boldsymbol{\theta}$ , we assign a  $\text{Normal}(\mathbf{0}, \sigma_\theta^2 \mathbf{I})$  prior, with variance large enough to remain noninformative.

Denote  $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2)^T$  and  $\boldsymbol{\phi} = (\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma_\epsilon^2)^T$ . Based on the above model and prior specification, the joint posterior distribution of all unknown parameters is

$$\begin{aligned} [\boldsymbol{\phi} | \mathbf{Y}] &\propto \prod_{\ell=0}^2 \gamma_\ell^{a_\ell + K/2 - 1} (\sigma_\epsilon^2)^{-(a_\ell + n/2) - 1} \exp\{-b_\epsilon / \sigma_\epsilon^2 - \sum_{\ell=0}^2 b_\ell \gamma_\ell - \boldsymbol{\theta}^T \boldsymbol{\theta} / (2\sigma_\theta^2)\} \\ &\quad \exp\{-\gamma_0 \boldsymbol{\zeta}^T(\boldsymbol{\beta}, \boldsymbol{\theta}) \boldsymbol{\zeta}(\boldsymbol{\beta}, \boldsymbol{\theta})/2 - \boldsymbol{\beta}^T(\gamma_1 H_1 + \gamma_2 H_2 + \gamma_1 \gamma_2 H_3) \boldsymbol{\beta}/2\} \end{aligned}$$



$$- (2\sigma_\epsilon^2)^{-1}(\mathbf{Y} - \mathbf{B}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{B}\boldsymbol{\beta})\}. \quad (3.15)$$

The posterior distribution (3.15) is not analytically tractable, hence we use a Markov chain Monte Carlo (MCMC) based computation method (Gilks, Richardson and Spiegelhalter, 1996) or more precisely Gibbs sampling (Gelfand and Smith, 1990) to simulate the parameters from the posterior distribution. To implement the Gibbs sampler, we need the full conditional distributions of all unknown parameters. Due to the choice of conjugate priors, the full conditional distributions of  $\sigma_\epsilon^2$  and  $\gamma_\ell$ 's are easily obtained as Inverse-Gamma and Gamma distributions, respectively. The full conditional distributions of  $\boldsymbol{\beta}$  and  $\boldsymbol{\theta}$  are not of standard form, and hence we employ Metropolis-Hastings algorithm to sample them.

In the special case of a linear PDE, simplifications arise. With approximation (3.5), the PDE modeling errors are represented as  $\zeta(\mathbf{x}_i) = \mathbf{f}^T(\mathbf{x}_i; \boldsymbol{\theta})\boldsymbol{\beta}$ , for  $i = 1, \dots, n$ . Define the matrix  $\mathbf{F}(\boldsymbol{\theta}) = \{\mathbf{f}(\mathbf{x}_1; \boldsymbol{\theta}), \dots, \mathbf{f}(\mathbf{x}_n; \boldsymbol{\theta})\}^T$ . Then the  $\boldsymbol{\beta}$  prior (3.14) becomes

$$[\boldsymbol{\beta} | \boldsymbol{\theta}, \gamma_0, \gamma_1, \gamma_2] \propto (\gamma_0 \gamma_1 \gamma_2)^{K/2} \exp[-\boldsymbol{\beta}^T \{\gamma_0 \mathbf{F}^T(\boldsymbol{\theta}) \mathbf{F}(\boldsymbol{\theta}) + \gamma_1 H_1 + \gamma_2 H_2 + \gamma_1 \gamma_2 H_3\} \boldsymbol{\beta} / 2],$$

where the exponent is quadratic in  $\boldsymbol{\beta}$ . And the posterior (3.15) becomes

$$\begin{aligned} [\boldsymbol{\phi} | \mathbf{Y}] &\propto \prod_{\ell=0}^2 \gamma_\ell^{a_\ell + K/2 - 1} (\sigma_\epsilon^2)^{-(a_\ell + n/2) - 1} \exp\{-b_\epsilon / \sigma_\epsilon^2 - \sum_{\ell=0}^2 b_\ell \gamma_\ell - \boldsymbol{\theta}^T \boldsymbol{\theta} / (2\sigma_\theta^2)\} \\ &\quad \exp[-\boldsymbol{\beta}^T \{\gamma_0 \mathbf{F}^T(\boldsymbol{\theta}) \mathbf{F}(\boldsymbol{\theta}) + \gamma_1 H_1 + \gamma_2 H_2 + \gamma_1 \gamma_2 H_3\} \boldsymbol{\beta} / 2 \\ &\quad - (2\sigma_\epsilon^2)^{-1}(\mathbf{Y} - \mathbf{B}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{B}\boldsymbol{\beta})]. \end{aligned}$$

Under linear PDE models, the full conditional of  $\boldsymbol{\beta}$  is easily seen to be a Normal distribution. This reduces the computational cost significantly compared with sampling under nonlinear cases, because the length of the vector  $\boldsymbol{\beta}$  increases quickly as dimension increases. Computational details of both nonlinear and linear PDE are shown in Appendix A.4.

### 3.4.1. Bayesian P-Splines

Here we describe briefly the implementation of Bayesian penalized splines, or P-splines. The term was first brought out by Eilers and Marx (1996). Also see Eilers and Marx (2010), Ruppert et al. (2003) for reference on univariate function smoothing. Eilers and Marx (2003), Marx and Eilers (2005), and Xiao, Li and Ruppert (2010) deal specifically with bivariate penalized B-splines. We use the bivariate B-spline basis, which is formed by tensor product of one-dimensional B-spline basis. Following Xiao et al. (2010) in our implementation, the difference penalty penalizes the interaction of one-dimensional coefficients as well as each dimension individually.

Denote the number of basis functions in each dimension by  $k_\ell$ , the one-dimensional basis function matrices by  $\mathbf{B}_\ell$ , and  $m_\ell^{th}$  order difference matrix of size  $(k_\ell - m_\ell) \times k_\ell$  by  $\mathbf{D}_\ell$ , for  $\ell = 1, 2$ . The prior density of the basis function coefficient  $\boldsymbol{\beta}$  of length  $K = k_1 k_2$  is assumed to be  $[\boldsymbol{\beta} | \gamma_1, \gamma_2] \propto (\gamma_1 \gamma_2)^{K/2} \exp\{-\boldsymbol{\beta}^T (\gamma_1 H_1 + \gamma_2 H_2 + \gamma_1 \gamma_2 H_3) \boldsymbol{\beta} / 2\}$ , where  $\gamma_1$  and  $\gamma_2$  are hyper-parameters, and the matrices are

$$H_1 = \mathbf{B}_1^T \mathbf{B}_1 \otimes \mathbf{D}_2^T \mathbf{D}_2; H_2 = \mathbf{D}_1^T \mathbf{D}_1 \otimes \mathbf{B}_2^T \mathbf{B}_2; H_3 = \mathbf{D}_1^T \mathbf{D}_1 \otimes \mathbf{D}_2^T \mathbf{D}_2. \quad (3.16)$$

When assuming conjugate prior distributions as  $[\sigma_\epsilon^2] = \text{IG}(a_\epsilon, b_\epsilon)$ ,  $[\gamma_1] = \text{Gamma}(a_1, b_1)$ , and  $[\gamma_2] = \text{Gamma}(a_2, b_2)$ , the posterior distribution can be derived easily and sampled using the Gibbs sampler. Though the prior distribution of  $\boldsymbol{\beta}$  is improper, the posterior distribution is proper (Berry, Carroll and Ruppert, 2002).

### 3.5. Simulations

In this section, the finite sample performances of the proposed parameter cascading and Bayesian method are investigated via Monte Carlo simulations, which are also compared with a two-stage method described below.

### 3.5.1. A Two-Stage Method

The two-stage method is constructed for PDE parameter estimation as follows. In the first stage,  $g(\mathbf{x})$  and the partial derivatives of  $g(\mathbf{x})$  are estimated by the Bayesian regression P-spline method described in Section 3.4.1. Let  $\hat{\boldsymbol{\beta}}$  denote the estimated coefficients of the basis functions in the first stage. In the second stage, we plug the estimated function and partial derivatives into the PDE model for each observation, i.e., we form  $\hat{\mathcal{F}}\{\hat{g}(\mathbf{x}_i); \boldsymbol{\theta}\}$  for  $i = 1, \dots, n$ . Then, a least-square type estimator for the PDE parameter,  $\boldsymbol{\theta}$ , is obtained by minimizing  $J(\boldsymbol{\theta}) = \sum_{i=1}^n \hat{\mathcal{F}}^2\{\hat{g}(\mathbf{x}_i); \boldsymbol{\theta}\} = \hat{\boldsymbol{\beta}}^T \{\sum_{i=1}^n \hat{\mathbf{f}}(\mathbf{x}_i; \boldsymbol{\theta}) \hat{\mathbf{f}}^T(\mathbf{x}_i; \boldsymbol{\theta})\} \hat{\boldsymbol{\beta}}$ , which is the sum of squared residuals of the fitted PDE model. For comparison purpose, the standard errors of two-stage estimates of the PDE parameters are estimated using parametric bootstrap.

### 3.5.2. Data Generating Mechanism

The PDE model (3.2) proposed for the LIDAR data set described in Section 3.6 is used to simulate data. The PDE model (3.2) is numerically solved using Matlab built-in facility by setting the true parameter values as  $\theta_D = 1, \theta_S = 0.1$ , and  $\theta_A = 0.1$ , the boundary condition as  $g(t, 0) = 0$ , and the initial condition as  $g(0, z) = \{1 + 0.1 \times (20 - z)^2\}^{-1}$  over a meshgrid in the time domain  $t \in [1, 20]$  and the range domain  $z \in [1, 40]$ . In order to obtain a precise numerical solution, we take grid of size 0.0005 in the time domain and size 0.001 in the range domain.

A sketch of the numerical solution is shown in Figure 5. In Figure 6, we show sketches of true partial derivatives, approximated simply by finite difference. Then the observed error-prone data is simulated by adding *i.i.d* Gaussian noise with  $\sigma^2 = 0.02^2$  to the PDE solutions at every 1 time unit and every 1 range unit, i.e., our data is on a 20-by-40 meshgrid in the domain  $[1, 20] \times [1, 40]$ .

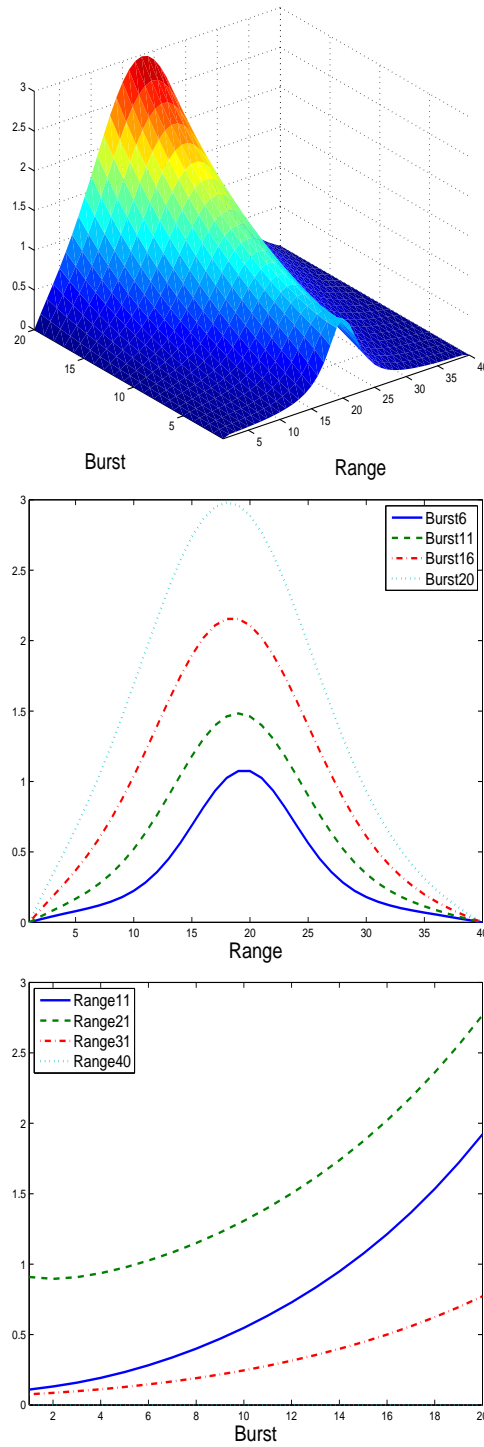


Figure 5. Snapshots of the solution function  $g(t, z)$ , i.e., the error-free data. Top: 3-D plot of the surface  $g(t, z)$ . Middle: plot of  $g(t_i, z)$  for time values  $t_i$  over range, with indices  $i = 1, 6, 11, 16, 20$ . Bottom: plot of  $g(t, z_j)$  for range values  $z_j$  over time, with indices  $j = 1, 11, 21, 31, 40$ .

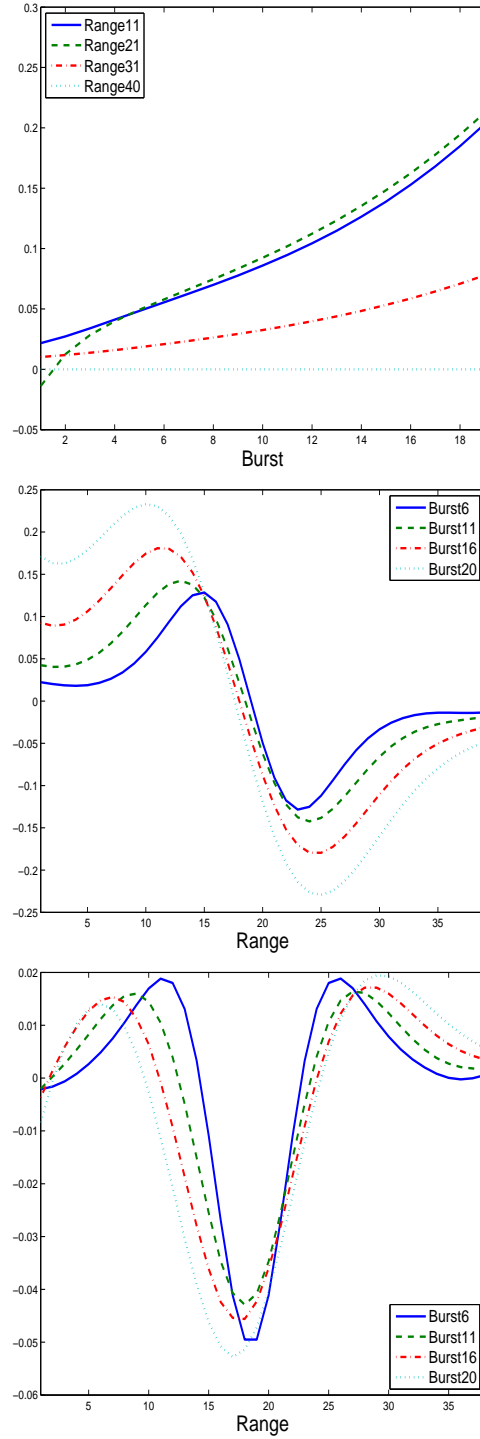


Figure 6. Snapshots of the partial derivatives of  $g(t, z)$ . Top: plot of  $\partial g(t, z_j)/\partial t$  for range values  $z_j$  over time, with indices  $j = 1, 11, 21, 31, 40$ . Middle: plot of  $\partial g(t_i, z)/\partial z$  for time values  $t_i$  over range with indices  $i = 1, 6, 11, 16, 20$ . Bottom: plot of  $\partial^2 g(t_i, z)/\partial z^2$  for time values  $t_i$  over range, with indices  $i = 1, 6, 11, 16, 20$ .

### 3.5.3. Performance of the Proposed Methods

The parameter cascading method, the Bayesian method, and the two-stage method were applied to estimate the three parameters in the PDE model (3.2). There were 1,000 simulated data sets. This section summarizes the performance of these three methods.

The PDE model (3.2) indicates that the second partial derivative with respect to  $z$  is continuously differentiable, and thus we choose quartic basis functions in the range domain. Therefore, in approximating the dynamic process  $g(t, z)$ , we use a tensor product of one-dimensional quartic B-splines to form the basis functions, with 5 and 17 equally spaced knots in time domain and range domain, respectively, in all three methods.

In the two-stage method for estimating PDE parameters, the Bayesian P-Splines method is used to estimate the dynamic process and the derivatives by setting the hyper-parameters defined in Section 3.1 as  $a_\epsilon = b_\epsilon = a_1 = b_1 = a_2 = b_2 = 0.01$ , and taking the third order difference matrix to penalize the roughness of the second derivative in each dimension. In the Bayesian method for estimating PDE parameters, we take the same smoothness penalty as in the two-stage method, and the hyper-parameters defined in Section 3.4 are set to be  $a_\epsilon = b_\epsilon = 0.01$ ,  $a_\ell = b_\ell = 0.001$  for  $\ell = 0, 1, 2$ , and  $\sigma_\theta^2 = 3^2$ . In the MCMC sampling procedure, we collect every 5<sup>th</sup> sample after a burn-in stage of length 5,000, until 3,000 posterior samples are obtained.

Table 3. The biases, standard deviations (SDs), square root of mean squared errors (RMSEs) of the parameter estimates for the PDE model (3.2) using the Bayesian method (BM), the parameter cascading method (PC), and the two-stage method (TS) in the 1000 simulation replicates. The coverage probabilities (CP) of on 95% credible/confidence intervals are also shown. The true parameter values are shown in the second row. As the two-stage method results in significant bias, we skip variance calculation for this method, and no coverage probability is provided.

		$\theta_D$	$\theta_S$	$\theta_A$
True		1	0.1	0.1
Bias	BM	-0.0165	-0.00042	-0.00016
	PC	-0.0297	-0.00013	-0.00027
	TS	-0.2252	-0.00068	-0.00183
SD	BM	$9.07 \times 10^{-3}$	$1.60 \times 10^{-3}$	$2.18 \times 10^{-4}$
	PC	$2.49 \times 10^{-2}$	$3.75 \times 10^{-3}$	$4.65 \times 10^{-4}$
	TS	$9.09 \times 10^{-2}$	$5.87 \times 10^{-3}$	$1.12 \times 10^{-3}$
RMSE	BM	$1.88 \times 10^{-2}$	$1.66 \times 10^{-3}$	$2.72 \times 10^{-4}$
	PC	$3.89 \times 10^{-2}$	$3.75 \times 10^{-3}$	$5.36 \times 10^{-4}$
	TS	$2.43 \times 10^{-1}$	$5.91 \times 10^{-3}$	$2.06 \times 10^{-2}$
CP	BM	0.939	0.999	0.988
	PC	0.775	0.946	0.915
	TS	N/A		

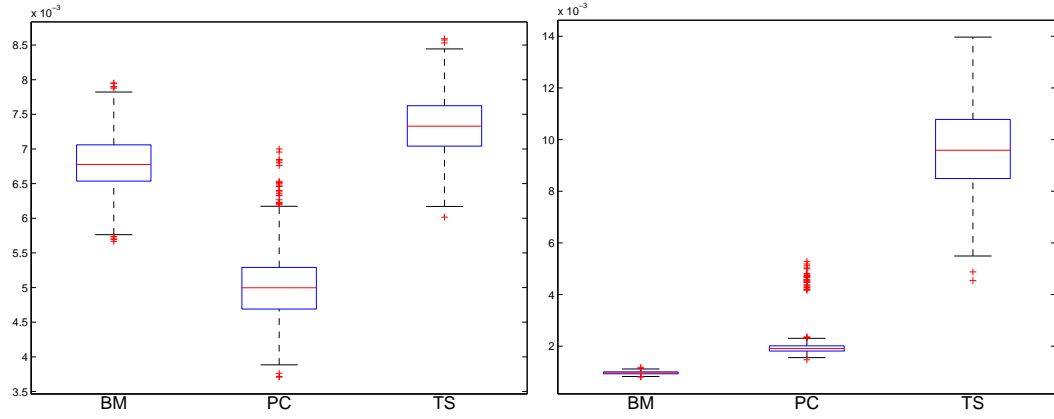


Figure 7. Boxplots of the square root of average squared errors (RASE) from 1000 data sets in the simulation study. Left: boxplots of  $\text{RASE}(\hat{g})$ , the solution function estimation, by all three methods. Right: boxplots of  $\text{RASE}(\hat{\mathcal{F}})$ , the PDE model estimation, by all three methods. The three methods produce similar data fitting, but the parameter cascading and Bayesian methods result in better PDE fitting.

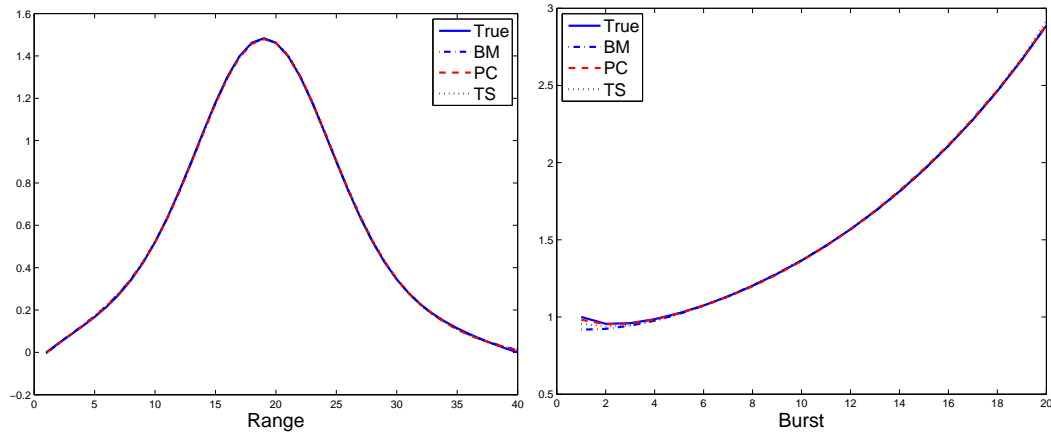


Figure 8. Cross-sectional views of the estimated solution in one data set in the simulation study. Left: function  $\hat{g}(t_{11}, z)$ . Right: function  $\hat{g}(t, z_{20})$ . All three methods produce similar function estimation.



To assess the accuracy of a bivariate function estimator  $\widehat{g}(t, z)$ , we use the square root of average squared error (RASE), which is defined as

$$\text{RASE}(\widehat{g}) = \left[ m_{\text{tgrid}}^{-1} m_{\text{zgrid}}^{-1} \sum_{j=1}^{m_{\text{tgrid}}} \sum_{k=1}^{m_{\text{zgrid}}} \{\widehat{g}(t_j, z_k) - g(t_j, z_k)\}^2 \right]^{1/2}, \quad (3.17)$$

where  $m_{\text{tgrid}}$  and  $m_{\text{zgrid}}$  are the number of grid points in each dimension, and  $t_j, z_k$  are grid points for  $j = 1, \dots, m_{\text{tgrid}}$ , and  $k = 1, \dots, m_{\text{zgrid}}$ .

We summarize the simulation results of the three estimators in the 1000 simulation replicates in Table 3, including the biases, standard deviations (SDs), square root of mean squared errors (RMSEs), and coverage probabilities of 95% confidence intervals for each method. We see that Bayesian method and parameter cascading method are comparable, and both have smaller biases, SDs and RMSEs than the two-stage method. Specifically, the improvement in the first parameter is substantial, which is associated with the second partial derivative,  $\partial^2 g(t, z)/\partial z^2$ . This is consistent with our conjecture that the two-stage strategy is not statistically efficient because of the inaccurate estimation of derivatives, especially higher order derivatives.

Figure 7 presents the boxplots of RASEs for the estimated dynamic process  $\widehat{g}$  and PDE model  $\widehat{\mathcal{F}}$  in the 1000 simulation replicates for each method. Again, we see improvement in function estimation, especially in fitting the PDE model. Taking an arbitrary simulated data set, we show in Figure 8 two cross-sectional views of the estimated solution surface  $\widehat{g}(t, z)$ , at burst 11 and range 30, respectively. Figure 9 shows cross-sectional views of the estimated partial derivatives. The estimated solutions by three methods almost overlap with each other, but there is a difference among the derivative estimation, especially second order derivative.

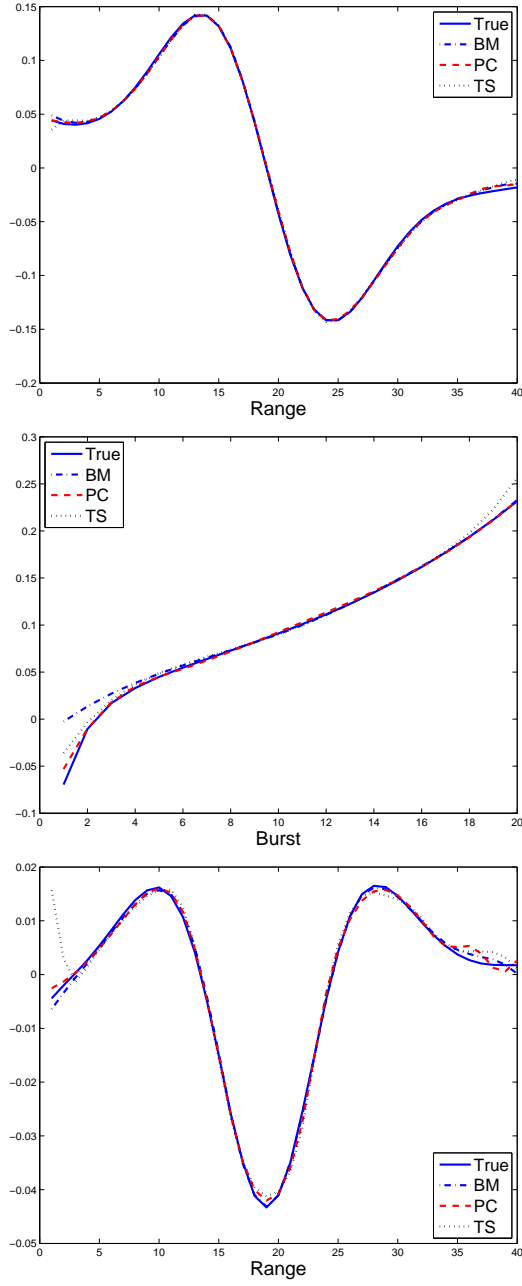


Figure 9. Estimated partial derivatives of the curves shown in Figure 8. Top: function  $\partial \hat{g}(t_{11}, z) / \partial z$ . Middle: function  $\partial \hat{g}(t, z_{20}) / \partial t$ . Bottom: function  $\partial^2 \hat{g}(t_{11}, z) / \partial z^2$ . There is a clear difference in the estimated second partial derivative between the two-stage method and the joint modeling methods.

### 3.6. Empirical Example

#### 3.6.1. The Example

We have access to data from an experiment involving long range infrared light detection and ranging (LIDAR) method. The goal of the experiment is to detect, locate and identify potentially hazardous aerosols, e.g., aerosols containing ovalbumin. Data were collected at a number of CO<sub>2</sub> laser wavelengths. At each wavelength, signals (or waveforms) were sent out every second for about 150 bursts. For each burst, received LIDAR data were observed at 625 equally spaced sampling points over the range. We propose to use the PDE (3.2) to model the data collected at a single wavelength. Simply put, this model describes, for example, how the initial concentration diffuses, shifts, and reacts to an additional force  $g(t, z)$  over time. The rate of diffusion, direction and rate of shift, and reaction to  $g(t, z)$  are reflected by  $\theta_D$ ,  $\theta_S$  and  $\theta_A$ , respectively.

In fitting model (3.2) to the real data, we only consider the middle 20 bursts and middle 60 range values, where the most information is contained. Burst values and range values are integers starting from 1. The sample size  $n$  is this  $20 \times 60 = 1,200$ . Snapshots of the data are shown in Figure 10.

#### 3.6.2. Results

The parameter cascading method, Bayesian method, and the two-stage method are applied to estimate the three parameters in the PDE model (3.2) from the above LIDAR data set. All three methods use bivariate quartic B-spline basis functions constructed with 5 inner knots in the burst domain and 20 inner knots in the range domain.

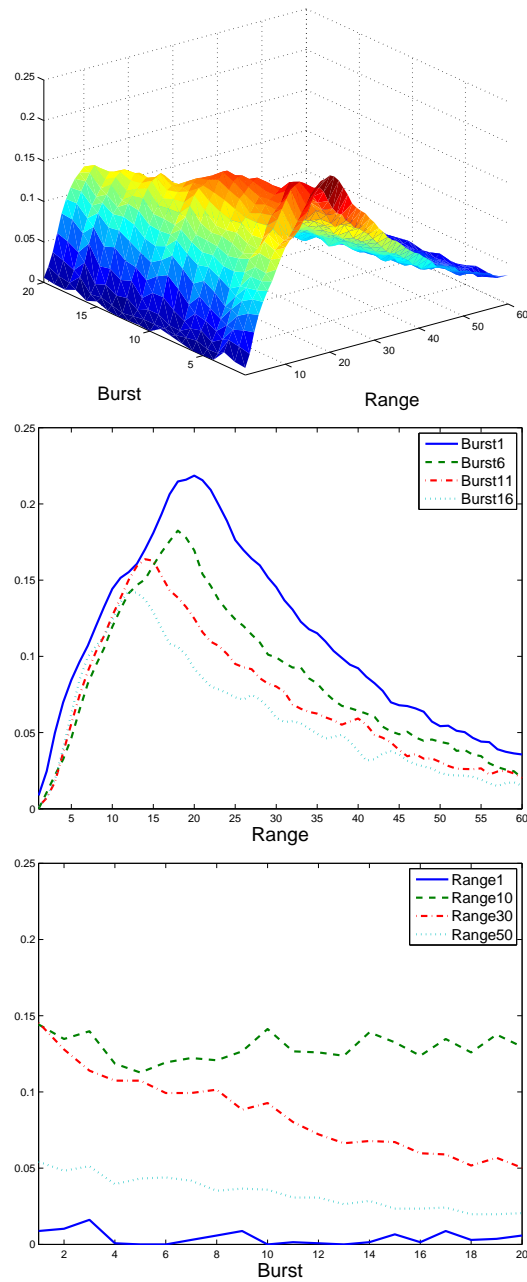


Figure 10. Snapshots of empirical data. Top: 3D plot of the received signal. Middle: the received signal at a few burst values over range. Bottom: the received signal at a few range values over burst.

Table 4. Results of the empirical example. We show the estimates for each parameter, and corresponding standard errors (SE). Methods: BM = Bayesian method; PC = parameter cascading method; TS = two-stage method.

		$\theta_D$	$\theta_S$	$\theta_A$
Estimates	BM	-0.4470	0.2563	-0.0414
	PC	-0.3771	0.2492	-0.0407
	TS	-0.1165	0.2404	-0.0436
SE	BM	$2.00 \times 10^{-1}$	$3.00 \times 10^{-2}$	$2.66 \times 10^{-3}$
	PC	$4.04 \times 10^{-3}$	$4.92 \times 10^{-5}$	$3.77 \times 10^{-7}$
	TS	N/A		

Table 5. Summary of function estimation in the empirical example. Estimated square root of average squared errors of  $g(t, z)$  and PDE  $\mathcal{F}(\cdot)$  estimation,  $\widehat{\text{RSAE}}(\hat{g})$  and  $\widehat{\text{RSAE}}(\hat{\mathcal{F}})$ , are shown. Methods: BM = Bayesian method; PC = parameter cascading method; TS = two-stage method.

	$\widehat{\text{RSAE}}(\hat{g})$	$\widehat{\text{RSAE}}(\hat{\mathcal{F}})$
BM	0.0051	0.0020
PC	0.0049	0.0010
TS	0.0046	0.0031

Table 4 displays the estimates for the three parameters in the PDE model (3.2) and their standard errors. While the three methods produce similar estimates for parameters  $\theta_S$  and  $\theta_A$ , the parameter cascading estimate and Bayesian estimate for  $\theta_D$  are more consistent with each other than with the two-stage estimate. This results is in accordance with the simulation study, where both the parameter cascading and Bayesian methods improve substantially the estimation of  $\theta_D$ , with comparable

estimates. The Bayesian standard errors are estimated posterior standard deviations, i.e., sample standard deviations of posterior samples. The parameter cascading standard errors are the estimates of the theoretical derivation provided in Section 3.3.3. The Bayesian method generally results in larger standard errors because the Bayesian modeling approach incorporates more uncertainty by averaging over many spline models according to a distribution of  $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2)$ . Whereas, the frequentist method estimates  $\boldsymbol{\theta}$  under a fixed spline model, with an optimal smoothing parameter  $\lambda$ .

We point out in Section 3.3.3 that the parameter cascading estimators are asymptotically Normal. With a Normal distribution approximation, it is obvious that  $\hat{\boldsymbol{\theta}}$  by the parameter cascading method is statistically significant. However, for the Bayesian method, the posterior samples do not necessarily follow a Normal distribution. We make inference using the sample 95% credible intervals (CI) of each parameter. The 95% CI's are  $(-0.3131, -0.0239)$ ,  $(0.1980, 0.3111)$  and  $(-0.0445, -0.0340)$  for  $\theta_D$ ,  $\theta_S$  and  $\theta_A$ , respectively. Hence the Bayesian method also concludes that the estimation is significant.

Function estimation is summarized in Table 5. We propose to use the RASE defined in (3.17) as a measure of solution function fit and PDE model fit. When analyzing real data, we replace the unknown solution values with observed data in the definition, and the RASE of  $\hat{g}$  is estimated as

$$\widehat{\text{RASE}}(\hat{g}) = \left[ m_{\text{tgrid}}^{-1} m_{\text{zgrid}}^{-1} \sum_{j=1}^{m_{\text{tgrid}}} \sum_{k=1}^{m_{\text{zgrid}}} \{\hat{g}(t_j, z_k) - Y_{jk}\}^2 \right]^{1/2},$$

where  $Y_{jk}$  is the observation at grid point  $(t_j, z_k)$ . Table 5 confirms that the three methods perform similarly in solution function estimation, and the new methods perform better in estimating the PDE model. For an intuitive understanding of the

fits, we also show some plots of the estimated function and its derivatives. Figure 11 displays the cross-sectional views of the estimated solution at burst index 11 and range index 30, and Figure 12 shows the cross-sectional views of estimated partial derivatives. As in the simulation, three methods produce almost identical smooth curves, but not the derivatives.

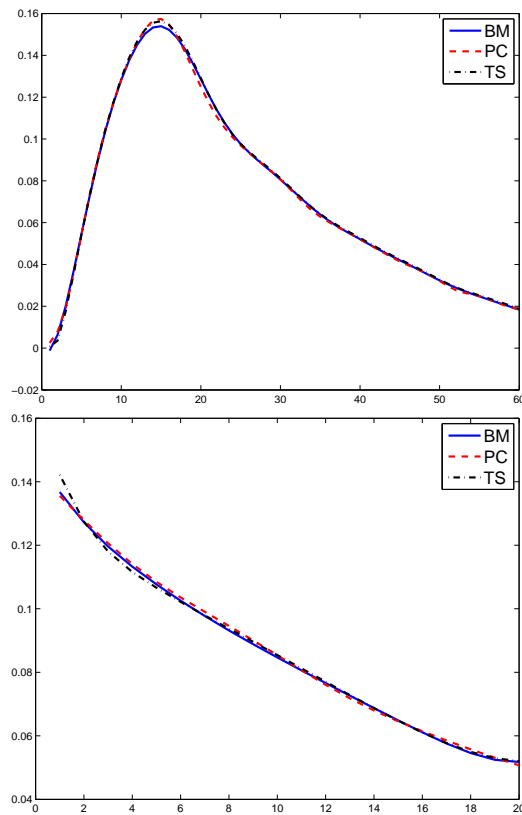


Figure 11. Cross-sectional views of estimated solution in the empirical example. Top: function  $\hat{g}(t_{11}, z)$ . Bottom: function  $\hat{g}(t, z_{30})$ . Three methods produce similar function estimation.

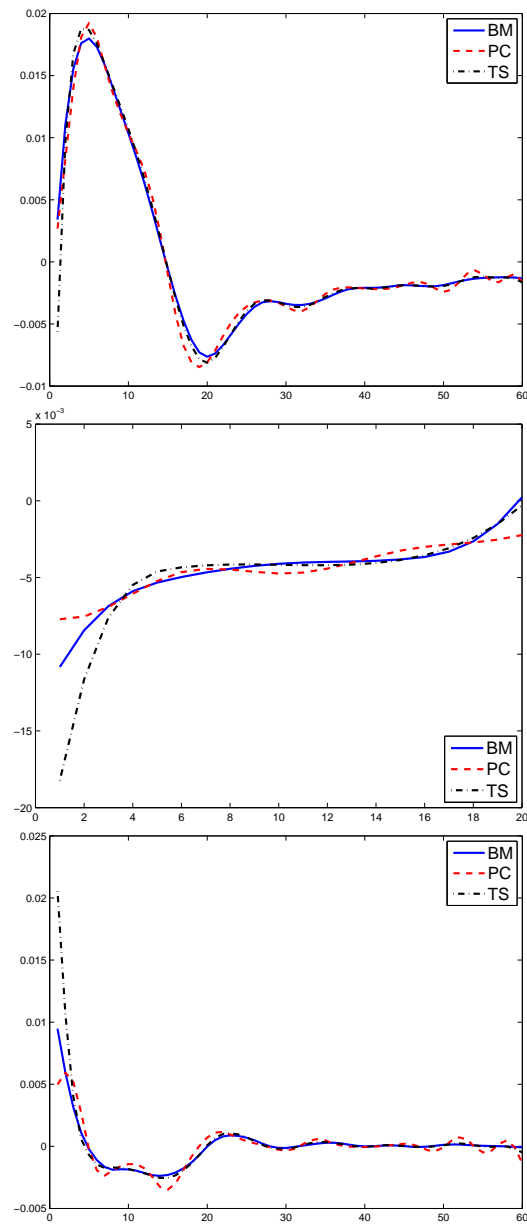


Figure 12. Cross-sectional views of estimated derivatives of curves shown in Figure 11. Top: function  $\partial \hat{g}(t_{11}, z) / \partial z$ . Middle: function  $\partial \hat{g}(t, z_{30}) / \partial t$ . Bottom: function  $\partial^2 \hat{g}(t_{11}, z) / \partial z^2$ .



### 3.7. Concluding Remarks

Differential equation (DE) models are widely used to model dynamic processes in many fields such as engineering and biomedical sciences. The forward problem of solving equations or simulating state variables for given parameters that define the DE models has been extensively studied in the past. However, the inverse problem of estimating parameters based on observed state variables is relatively sparse in the statistical literature, and this is especially the case for partial differential equation (PDE) models.

We propose a parameter cascading method and a fully Bayesian treatment for this problem, and compare them with a straightforward two-stage method. The parameter cascading method and Bayesian method are joint estimation procedures which consider the data fitting and PDE fitting simultaneously. Hence the proposed methods are more statistically efficient than a two-stage method, as confirmed by the simulation studies. The improvement is substantial for parameters associated with higher order derivatives. Basis function expansion plays an important role in our new methods, in the sense that it makes joint modeling possible and links together fidelity to PDE model and fidelity to data through the coefficients of basis functions. A potential extension of this work would be to estimate functional parameters in PDEs from error-prone data.

Simulation studies show that the parameter cascading method and Bayesian method provides comparable parameter estimates for PDE models, which are accurate than the two-stage method. We also apply the proposed methods to an empirical LIDAR data set, which is from a remote source detection problem. In this example, we propose to use PDE model (3.2) to fit the received signals.

## CHAPTER IV

### CONCLUSION

Inverse problems arise widely in many different fields, from science like physics and biology to engineering like medical imaging and remote sensing. They are gaining popularity in statistical society in recent years. From a statistical point of view, inverse problems are recast as problems of statistical inference. Bayesian statistics is widely used in this area. As a unified modeling framework, problems are solved in a systematical way. It also has the advantage of naturally and properly incorporating all available information into model. Whereas, frequentist solutions are problem specific.

In the source detection problem, we consider the problem of detecting existence of a low emission radiating source inside a volume, in the presence of a strong random background. We are interested in the situation when only about 1% of detected hits are by the ballistic particles coming from the source. We treat it as a model selection problem and develop a Bayesian approach. Decision are made based on the collected data and the value of the corresponding Bayes factors, which model fits better the collected data. A simulation study shows that Bayesian methods can be successfully used for detection of low emission small sources in the cases of realistic parameters. High sensitivity can be achieved if the observation time is sufficiently large.

The assumption that the detector can determine the directional information is crucial. We believe that otherwise detection with such low values of signal-to-noise ratio would be impossible. In practice, the so called Compton type cameras for detecting  $\gamma$ -photons (e.g., Allmaras et al., 2010, LeBlanc et al., 1998 and references therein), as well as their analogs (although based upon a somewhat different physics) for neutron detection (Marianno et al., 2010; Spence and Charlton, 2009) could be used to collect data. These cameras can determine a hollow cone of possible directions

of the incoming particle. Although this is a less precise and highly over-determined information, it is known (e.g., see Allmaras et al., 2010) how to convert the Compton type data into the precise directional information. Thus, the algorithms described can be used in conjunction with Compton type detectors.

Differential equation models are widely used to model dynamic processes in many fields such as engineering and biomedical sciences. The forward problem of solving equations or simulating state variables for given parameters that define the models have been extensively studied in the past. However, the inverse problem of estimating parameters based on observed state variables is relatively sparse in the statistical literature, and this is especially the case for partial differential equation models.

We have proposed a parameter cascading method and a fully Bayesian treatment for this problem, and compared them with a straightforward two-stage method. The parameter cascading method and Bayesian method are joint estimation procedures which consider the data fitting and PDE fitting simultaneously. Hence the proposed methods are more statistically efficient than a two-stage method, as confirmed by the simulation studies. The improvement is substantial for parameters associated with higher order derivatives. Basis function expansion plays an important role in our new methods, in the sense that it makes joint modeling possible and links together fidelity to the PDE model and fidelity to data through the coefficients of basis functions. A potential extension of this work would be to estimate functional parameters in PDE from error-prone data.

## REFERENCES

- Allmaras, M., Darrow, D., Hristova, Y., Kanschat, G., and Kuchment, P. (2010). Detecting small low emission radiating sources. Preprint. Online version: arXiv:1012.3373.
- Bar, M., Hegger, R., and Kantz, H. (1999). Fitting differential equations to space-time dynamics. *Physical Review E* **59**, 337-342.
- Berry, S. M., Carroll, R. J., and Ruppert, D. (2002). *Journal of the American Statistical Association* **97**, 160-169.
- Brenner, S. C. and Scott, R. (2010). *The Mathematical Theory of Finite Element Methods*. New York: Springer.
- Budinger, T. F., Gullberg, G. T., and Huesman R. H. (1979). Emission computed tomography. *Image Reconstruction from Projections: Implementation and Applications. Volumn 32 in Topics in Applied Physics*. Herman, G. T. (ed.) New York: Springer.
- Burden, R. L. and Douglas, F. J. (2010). *Numerical Analysis*, Ninth edition. California: Brooks/ Cole.
- Cao, J., Huang, J. Z., and Wu, H. (2011). Penalized nonlinear least squares estimation of time-varying parameters in ordinary differential equations. *Journal of Computational and Graphical Statistics*, to appear. Online version DOI: 10.1198/jcgs.2011.10021.
- Cao, J., Wang, L., and Xu, J. (2011). Robust estimation for ordinary differential equation models. *Biometrics* **67**, 1305-1313.

- Chen, J. and Wu, H. (2008). Efficient local estimation for time-varying coefficients in deterministic dynamic models with applications to HIV-1 dynamics. *Journal of the American Statistical Association* **103**, 369-384.
- Crainiceanu, C. M. and Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society, Series B* **66**, 165-185.
- de Boor, C. (2001). *A Practical Guide to Splines*. Revised edition. Applied Mathematical Sciences 27. New York: Springer.
- Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1997). Automatic Bayesian curve fitting. *Journal of the Royal Statistical Society, Series B* **60**, 333-350.
- Eilers, P. and Marx, B. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11(2)**, 89-121.
- Eilers, P. and Marx, B. (2003). Multidimensional calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and Intelligent Lab Systems* **66**, 159-174.
- Eilers, P. and Marx, B. (2010). Splines, knots and penalties. *Wiley Interdisciplinary Reviews: Computational Statistics* **2**, 637-653.
- Evans, L. C. (1998). *Partial Differential Equations*. Graduate Studies in Mathematics 19. American Mathematical Society.
- Evett, I. W. (1991). Implementing Bayesian methods in forensic science. Paper presented at the Fourth Valencia International Meeting on Bayesian Statistics, Valencia, Spain.
- Friedman, J. H. and Silverman, B. W. (1989). Flexible parsimonious smoothing and

- additive modeling. *Technometrics* **31**, 3-21.
- Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398-409.
- Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice: Interdisciplinary Statistics*. London: Chapman & Hall.
- Good, I. J. (1985). Weight of evidence: a brief survey. *Bayesian Statistics 2*. Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M. (eds.) New York: Elsevier.
- Goswami, G. and Liu J. S. (2007). On learning strategies for evolutionary Monte Carlo. *Statistics and Computing* **17**, 23-28.
- Ho, D. D., Neumann, A. S., Perelson, A. S., Chen, W., Leonard, J. M., and Markowitz, M. (1995). Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature* **373**, 123-126.
- Huang, Y., Liu, D., and Wu, H. (2006). Hierarchical Bayesian methods for estimation of parameters in a longitudinal HIV dynamic system. *Biometrics* **62**, 413-423.
- Huang, Y. and Wu, H. (2006). A Bayesian approach for estimating antiviral efficacy in HIV dynamic models. *Journal of Applied Statistics* **33**, 155-174.
- Jeffreys, H. (1961). *Theory of Probability*, 3rd edition. Oxford: Oxford University Press.
- Kaipio, J. and Somersalo, E. (2005). *Statistical and Computational Inverse Problems*. New York: Springer.
- Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773-795.

- LeBlanc, J. W., Clinthorne, N. H., Hua, C. H., Nygard, E., Rogers, W. L., Wehe, D. K., Weilhammer, P., and Wilderman, S. J. (1998). C-SPRINT: a prototype compton camera system for low energy gamma-ray imaging. *IEEE Transactions on Nuclear Science* **45**, 943-949.
- Li, L., Brown, M. B., Lee, K. H., and Gupta, S. (2002). Estimation and inference for a spline-enhanced population pharmacokinetic model. *Biometrics* **58**, 601-611.
- Liang, F., Liu, C., and Carroll, R.J (2010). *Advanced Markov Chain Monte Carlo: learning from past samples*. New York: Wiley.
- Liang, H. and Wu, H. (2008). Parameter estimation for differential equation models using a framework of measurement error in regression models. *Journal of the American Statistical Association* **103**, 1570-1583.
- Marianno, C. M., Boyle, D. R., Charlton, W. S., and Gaukler, G. M. (2010). A guide for detector development and deployment. *Proceedings of the 2010 Annual Meeting of the Institute of Nuclear Materials Management*, Baltimore, Maryland.
- Marx, B. and Eilers, P. (2005). Multidimensional penalized signal regression. *Technometrics* **47**, 13-22.
- Morton, K. W. and Mayers, D. F. (2005). *Numerical Solution of Partial Differential Equations, An Introduction*. Cambridge: Cambridge University Press.
- Muller, T. and Timmer, J. (2002). Fitting parameters in partial differential equations from partially observed noisy data. *Physical Review D* **171**, 1-7.
- Muller, T. and Timmer, J. (2004). Parameter identification techniques for partial differential equations. *International Journal of Bifurcation and Chaos* **14**, 2053-2060.

- Parlitz, U. and Merkwirth, C. (2000). Prediction of spatiotemporal time series based on reconstructed local states. *Physical Review Letters* **84**, 1890-1893.
- Poyton, A. A., Varziri, M. S., McAuley, K. B., McLellan, P. J., and Ramsay, J. O. (2006). Parameter estimation in continuous-time dynamic models using principal differential analysis. *Computer and Chemical Engineering* **30**, 698-708.
- Putter, H., Heisterkamp, S. H., Lange, J. M. A., and De Wolf, F. (2002). A Bayesian approach to parameter estimation in HIV dynamical models. *Statistics in Medicine* **21**, 2199-2214.
- Qi, X. and Zhao, H. (2010). Asymptotic efficiency and finite-sample properties of the generalized profiling estimation of parameters in ordinary differential equations. *The Annals of Statistics* **38**, 435-481.
- Raftery, A. E. (1996). Hypothesis testing and model selection. *Markov Chain Monte Carlo in Practice: Interdisciplinary Statistics*. Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (eds.) London: Chapman & Hall.
- Ramsay, J. O. (1996). Principal differential analysis: data reduction by differential operators. *Journal of the Royal Statistical Society, Series B* **58**, 495-508.
- Ramsay, J. O., Hooker, G., Campbell, D., and Cao, J. (2007). Parameter estimation for differential equations: a generalized smoothing approach (with discussion). *Journal of the Royal Statistical Society, Series B* **69**, 741-796.
- Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge: Cambridge University Press.
- Spence, G. R. and Charlton, W. S. (2009). Directionally sensitive neutron detectors for portal monitors. *Proceedings of the 2009 Annual Meeting of the Institute of Nuclear Materials Management*, Tucson, Arizona.



- Stone, C. J., Hansen, M. H., Kooperberg, C., and Truong, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling. *Annals of Statistics* **25**, 1371-1425.
- Voss, H. U., Kolodner, P., Abel, M., and Kurths, J. (1999). Amplitude equations from spatiotemporal binary-fluid convection data. *Physical Review Letters* **83**, 3422-3425.
- Wand, M. P. (2000). A comparison of regression spline smoothing procedures. *Computational Statistics* **15**, 443-62.
- Wei, X., Ghosh, S. K., Taylor, M. E., Johnson, V. A., Emini, E. A., Deutsch, P., Lifson, J. D., Bonhoeer, S., Nowak, M. A., Hahn, B. H., Saag, M. S., and Shaw, G.M. (1995). Viral dynamics in human immunodeficiency virus type 1 infection. *Nature* **373**, 117-123.
- Wu, H. (2005). Statistical methods for HIV dynamic studies in AIDS clinical trials. *Statistical Methods in Medical Research* **14**, 171-192.
- Wu, H. and Ding, A. (1999). Population HIV-1 dynamics in vivo: applicable models and inferential tools for virological data from AIDS clinical trials. *Biometric* **55**, 410-418.
- Wu, H., Ding, A., and DeGruttola, V. (1998). Estimation of HIV dynamic parameters. *Statistics in Medicine* **17**, 2463-2485.
- Xiao, L., Li, Y., and Ruppert, D. (2010). Bivariate penalized splines. arXiv:1011.4916v1.
- Xun, X., Mallick, B. K., Carroll, R. J., and Kuchment, P. (2011). A Bayesian approach to the detection of small low emission sources. *Inverse Problems* **27**, 155009. Online version DOI: 10.1088/0266-5611/27/11/115009.

## APPENDIX A

### TECHNICAL DETAILS OF CHAPTER III

#### A.1 Calculation of Model Quantities

Here we show the form of  $\mathbf{f}(\mathbf{x}; \boldsymbol{\theta})$  and  $\mathbf{F}(\boldsymbol{\theta})$  for the PDE example (3.2). The vector  $\mathbf{f}(\mathbf{x}; \boldsymbol{\theta})$  is a linear combination of basis functions and their derivatives involved in model (3.2). We have that  $\mathbf{f}(\mathbf{x}; \boldsymbol{\theta}) = \partial \mathbf{b}(\mathbf{x}) / \partial t - \theta_D \partial^2 \mathbf{b}(\mathbf{x}) / \partial z^2 - \theta_S \partial \mathbf{b}(\mathbf{x}) / \partial z - \theta_A \mathbf{b}(\mathbf{x})$ . Similar to the basis function matrix  $\mathbf{B} = \{\mathbf{b}(\mathbf{x}_1), \dots, \mathbf{b}(\mathbf{x}_n)\}^T$ , we define the following  $n \times K$  matrices consisting of derivatives of the basis functions

$$\begin{aligned} \mathbf{B}_t &= \left\{ \frac{\partial \mathbf{b}(\mathbf{x}_1)}{\partial t}, \dots, \frac{\partial \mathbf{b}(\mathbf{x}_n)}{\partial t} \right\}^T, \\ \mathbf{B}_z &= \left\{ \frac{\partial \mathbf{b}(\mathbf{x}_1)}{\partial z}, \dots, \frac{\partial \mathbf{b}(\mathbf{x}_n)}{\partial z} \right\}^T, \\ \mathbf{B}_{zz} &= \left\{ \frac{\partial^2 \mathbf{b}(\mathbf{x}_1)}{\partial z^2}, \dots, \frac{\partial^2 \mathbf{b}(\mathbf{x}_n)}{\partial z^2} \right\}^T. \end{aligned}$$

Then the matrix  $\mathbf{F}(\boldsymbol{\theta}) = \{\mathbf{f}(\mathbf{x}_1; \boldsymbol{\theta}), \dots, \mathbf{f}(\mathbf{x}_n; \boldsymbol{\theta})\}^T = \mathbf{B}_t - \theta_D \mathbf{B}_{zz} - \theta_S \mathbf{B}_z - \theta_A \mathbf{B}$ .

#### A.2 Calculation of the Penalty Matrix

We have that  $\mathbf{R}(\boldsymbol{\theta})$  is a  $K \times K$  matrix which has  $(j, \ell)$  entry  $\int f_j(\mathbf{x}; \boldsymbol{\theta}) f_\ell(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x}$ . Using the notation of matrix integration, we write  $\mathbf{R}(\boldsymbol{\theta}) = \int \mathbf{f}(\mathbf{x}; \boldsymbol{\theta}) \mathbf{f}^T(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x}$ , where  $\mathbf{f}(\mathbf{x}; \boldsymbol{\theta}) = [f_1(\mathbf{x}; \boldsymbol{\theta}), \dots, f_K(\mathbf{x}; \boldsymbol{\theta})]^T$ . In our empirical work and simulations based on the PDE model (3.2), the penalty matrix  $\mathbf{R}(\boldsymbol{\theta})$  is the summation of 10 matrix integrals of the same structure, defined as

$$\mathbf{R}(\boldsymbol{\theta}) = \int_z \int_t \mathbf{f}(t, z; \boldsymbol{\theta}) \mathbf{f}^T(t, z; \boldsymbol{\theta}) dt dz$$

$$\begin{aligned}
&= \theta_D^2 \int \frac{\partial^2 \mathbf{b}}{\partial z^2} \frac{\partial^2 \mathbf{b}^T}{\partial z^2} d\mathbf{x} + \theta_S^2 \int \frac{\partial \mathbf{b}}{\partial z} \frac{\partial \mathbf{b}^T}{\partial z} d\mathbf{x} + \theta_A^2 \int \mathbf{b} \mathbf{b}^T d\mathbf{x} + \int \frac{\partial \mathbf{b}}{\partial t} \frac{\partial \mathbf{b}^T}{\partial t} d\mathbf{x} \\
&\quad + \theta_D \theta_S \int \left( \frac{\partial^2 \mathbf{b}}{\partial z^2} \frac{\partial \mathbf{b}^T}{\partial z} + \frac{\partial \mathbf{b}}{\partial z} \frac{\partial^2 \mathbf{b}^T}{\partial z^2} \right) d\mathbf{x} + \theta_D \theta_A \int \left( \frac{\partial^2 \mathbf{b}}{\partial z^2} \mathbf{b}^T + \mathbf{b} \frac{\partial^2 \mathbf{b}^T}{\partial z^2} \right) d\mathbf{x} \\
&\quad + \theta_S \theta_A \int \left( \frac{\partial \mathbf{b}}{\partial z} \mathbf{b}^T + \mathbf{b} \frac{\partial \mathbf{b}^T}{\partial z} \right) d\mathbf{x} - \theta_D \int \left( \frac{\partial^2 \mathbf{b}}{\partial z^2} \frac{\partial \mathbf{b}^T}{\partial t} + \frac{\partial \mathbf{b}}{\partial t} \frac{\partial^2 \mathbf{b}^T}{\partial z^2} \right) d\mathbf{x} \\
&\quad - \theta_S \int \left( \frac{\partial \mathbf{b}}{\partial z} \frac{\partial \mathbf{b}^T}{\partial t} + \frac{\partial \mathbf{b}}{\partial t} \frac{\partial \mathbf{b}^T}{\partial z} \right) d\mathbf{x} - \theta_A \int \left( \frac{\partial \mathbf{b}}{\partial t} \mathbf{b}^T + \mathbf{b} \frac{\partial \mathbf{b}^T}{\partial t} \right) d\mathbf{x} \\
&\triangleq \sum_{\ell=1}^L r_\ell(\boldsymbol{\theta}) \mathcal{B}_\ell, \tag{A.1}
\end{aligned}$$

where  $L = 10$ ,  $\mathcal{B}_\ell$  are known constant matrices, and  $r_\ell(\boldsymbol{\theta})$  are known functions of  $\boldsymbol{\theta}$ .

We compute  $\mathcal{B}_\ell$  for  $\ell = 1, \dots, 10$  following the same rule. In general, we can use the composite Simpson's rule repeatedly to evaluate the integrals. For a univariate function  $\phi(x)$  and an even integer  $Q$ , the composite Simpson's rule approximates the integral as

$$\begin{aligned}
\int_a^b \phi(x) dx &\approx (h/3) \left\{ \phi(x_0) + 2 \sum_{q=1}^{Q/2-1} \phi(x_{2q}) + 4 \sum_{q=1}^{Q/2} \phi(x_{2q-1}) + \phi(x_Q) \right\} \\
&= (h/3) \sum_{q=0}^Q w_q \phi(x_q),
\end{aligned}$$

where  $h = (b - a)/Q$ ,  $x_q = a + qh$ , for  $q = 0, 1, \dots, Q$ , are quadrature points, and  $(w_0, w_1, w_2, \dots, w_{Q-2}, w_{Q-1}, w_Q) = (1, 4, 2, \dots, 4, 2, 1)$  assigns weights to quadrature points.

In order to calculate, for example  $\mathcal{B}_3$ , let  $Q_1$  denote the number of quadrature knots in the time domain,  $\mathbf{s}_1 = (t_1, \dots, t_{Q_1})$  the vector of knots, and  $\mathbf{w}_1$  the vector of weights. Similarly,  $Q_2$ ,  $\mathbf{s}_2 = (z_1, \dots, z_{Q_2})$  and  $\mathbf{w}_2$  are the number of quadrature knots, knot vector, and weight vector in the range domain. Then the  $(i, j)$  entry  $\mathcal{B}_{3,ij}$  is

$$\mathcal{B}_{3,ij} = \int \int b_i(t, z) b_j(t, z) dt dz \approx (h/3)^2 \sum_{k=1}^{Q_2} \sum_{\ell=1}^{Q_1} w_{1,\ell} w_{2,k} b_i(t_\ell, z_k) b_j(t_\ell, z_k). \tag{A.2}$$

Define  $\mathbf{W}$  as a diagonal matrix with diagonal elements  $w_1 \otimes w_2$ . Denote the quadrature

points by  $\mathbf{Z} = \{(t_1, z_1), \dots, (t_1 z_{Q_2}), \dots, (t_{Q_1} z_1), \dots, (t_{Q_1} z_{Q_2})\}^T$ , and  $\mathbf{B}(\mathbf{Z})$  the matrix of basis function evaluated at the quadrature points. Then the approximation of matrix  $\mathcal{B}_3$  can be expressed neatly as,

$$\mathcal{B}_3 \approx \mathbf{B}^T(\mathbf{Z})\mathbf{W}\mathbf{B}(\mathbf{Z}). \quad (\text{A.3})$$

### A.3 Derivation of Variance Estimation in Section 3.3.3

We make the convention that all calculations assume that the  $\mathbf{x}_i$  are fixed. We also fix  $\tilde{\lambda} = \lambda/n$ , and consider  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)^T$  to be a  $m$ -dimensional parameter. The data model is

$$Y_i = g(\mathbf{x}_i) + \epsilon(\mathbf{x}_i),$$

where the  $\epsilon(\mathbf{x}_i)$  are independent of  $\mathbf{x}_i$  and have mean zero and variance  $\sigma_\epsilon^2$ . We use B-spline basis functions  $\mathbf{b}(\mathbf{x}_i) = \{b_1(\mathbf{x}_i), \dots, b_K(\mathbf{x}_i)\}^T$  to approximate  $g(\mathbf{x}_i) \approx \mathbf{b}^T(\mathbf{x}_i)\boldsymbol{\beta}$ . For a matrix  $\mathbf{R}(\boldsymbol{\theta})$  we define

$$\begin{aligned} \mathbf{S}_n &= n^{-1} \sum_{i=1}^n \mathbf{b}(\mathbf{x}_i) \mathbf{b}^T(\mathbf{x}_i); \\ \mathbf{G}_n(\boldsymbol{\theta}) &= \mathbf{S}_n + \tilde{\lambda} \mathbf{R}(\boldsymbol{\theta}); \\ \hat{\boldsymbol{\beta}}_n(\boldsymbol{\theta}) &= \mathbf{G}_n^{-1}(\boldsymbol{\theta}) n^{-1} \sum_{i=1}^n \mathbf{b}(\mathbf{x}_i) Y_i; \\ \boldsymbol{\beta}_n(\boldsymbol{\theta}) &= \mathbf{G}_n^{-1}(\boldsymbol{\theta}) n^{-1} \sum_{i=1}^n \mathbf{b}(\mathbf{x}_i) g(\mathbf{x}_i); \\ \mathbf{R}_{j\theta}(\boldsymbol{\theta}) &= \frac{\partial \mathbf{R}(\boldsymbol{\theta})}{\partial \theta_j}; \\ \boldsymbol{\Omega}_1 &= \lim_{n \rightarrow \infty} \mathbf{S}_n; \\ \boldsymbol{\Omega}_2(\boldsymbol{\theta}) &= \boldsymbol{\Omega}_1 + \tilde{\lambda} \mathbf{R}(\boldsymbol{\theta}). \end{aligned}$$

The parameter  $\boldsymbol{\theta}$  is estimated by minimizing

$$\mathcal{L}_n(\boldsymbol{\theta}) = n^{-1} \sum_{i=1}^n \{Y_i - \mathbf{b}^T(\mathbf{x}_i) \hat{\boldsymbol{\beta}}_n(\boldsymbol{\theta})\}^2.$$

### A.3.1 Calculating the Score and its Derivative

We remember the matrix fact that for any nonsingular symmetric matrix  $\mathbf{A}(z)$  for scalar  $z$ ,

$$\frac{\partial \mathbf{A}^{-1}(z)}{\partial z} = -\mathbf{A}^{-1}(z) \frac{\partial \mathbf{A}(z)}{\partial z} \mathbf{A}^{-1}(z).$$

This means that for  $j = 1, \dots, m$ ,

$$\begin{aligned} \frac{\partial \hat{\boldsymbol{\beta}}_n(\boldsymbol{\theta})}{\partial \theta_j} &= -\tilde{\lambda} \mathbf{G}_n^{-1}(\boldsymbol{\theta}) \mathbf{R}_{j\theta}(\boldsymbol{\theta}) \mathbf{G}_n^{-1}(\boldsymbol{\theta}) n^{-1} \sum_{i=1}^n \mathbf{b}(\mathbf{x}_i) Y_i \\ &= -\tilde{\lambda} \mathbf{G}_n^{-1}(\boldsymbol{\theta}) \mathbf{R}_{j\theta}(\boldsymbol{\theta}) \hat{\boldsymbol{\beta}}_n(\boldsymbol{\theta}). \end{aligned} \quad (\text{A.4})$$

Minimizing  $\mathcal{L}_n(\boldsymbol{\theta})$  is equivalent to solving for the system of equations

$$0 = n^{-1/2} \sum_{i=1}^n \{Y_i - \mathbf{b}^T(\mathbf{x}_i) \hat{\boldsymbol{\beta}}_n(\boldsymbol{\theta})\} \mathbf{b}^T(\mathbf{x}_i) \frac{\partial \hat{\boldsymbol{\beta}}_n(\boldsymbol{\theta})}{\partial \theta_j} = n^{-1/2} \sum_{i=1}^n \Psi_{ij}(\boldsymbol{\theta}), \quad j = 1, \dots, m,$$

where we define

$$\Psi_{ij}(\boldsymbol{\theta}) = \{Y_i - \mathbf{b}^T(\mathbf{x}_i) \hat{\boldsymbol{\beta}}_n(\boldsymbol{\theta})\} \mathbf{b}^T(\mathbf{x}_i) \frac{\partial \hat{\boldsymbol{\beta}}_n(\boldsymbol{\theta})}{\partial \theta_j}.$$

From now on, we define the score for  $\theta_j$  as

$$\mathcal{S}_{nj}(\boldsymbol{\theta}) = n^{-1/2} \sum_{i=1}^n \Psi_{ij}(\boldsymbol{\theta}) \quad (\text{A.5})$$

and define  $\mathcal{S}_n(\boldsymbol{\theta}) = \{\mathcal{S}_{n1}(\boldsymbol{\theta}), \dots, \mathcal{S}_{nm}(\boldsymbol{\theta})\}^T$ . Suppose the Hessian matrix is  $\mathcal{M}_n(\boldsymbol{\theta}) = \partial \mathcal{S}_n(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}^T$  so that the  $(j, k)^{th}$  element of  $\mathcal{M}_n(\boldsymbol{\theta})$  is

$$\mathcal{M}_{n,jk}(\boldsymbol{\theta}) = -n^{-1/2} \sum_{i=1}^n \frac{\partial \hat{\boldsymbol{\beta}}_n^T(\boldsymbol{\theta})}{\partial \theta_j} \mathbf{b}(\mathbf{x}_i) \mathbf{b}^T(\mathbf{x}_i) \frac{\partial \hat{\boldsymbol{\beta}}_n(\boldsymbol{\theta})}{\partial \theta_k}$$

$$\begin{aligned}
& +n^{-1/2} \sum_{i=1}^n \{Y_i - \mathbf{b}^T(\mathbf{x}_i) \hat{\boldsymbol{\beta}}_n(\boldsymbol{\theta})\} \mathbf{b}^T(\mathbf{x}_i) \frac{\partial^2 \hat{\boldsymbol{\beta}}_n(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k} \\
& = \mathcal{M}_{n1,jk}(\boldsymbol{\theta}) + \mathcal{M}_{n2,jk}(\boldsymbol{\theta}).
\end{aligned}$$

There are some further simplifications of  $\mathcal{S}_n(\boldsymbol{\theta})$ . Because of (A.4),

$$\mathcal{S}_{nj}(\boldsymbol{\theta}) = -\tilde{\lambda} n^{-1/2} \sum_{i=1}^n \{Y_i - \mathbf{b}^T(\mathbf{x}_i) \hat{\boldsymbol{\beta}}_n(\boldsymbol{\theta})\} \mathbf{b}^T(\mathbf{x}_i) \mathbf{G}_n^{-1}(\boldsymbol{\theta}) \mathbf{R}_{j\theta}(\boldsymbol{\theta}) \hat{\boldsymbol{\beta}}_n(\boldsymbol{\theta}).$$

However,

$$\begin{aligned}
n^{-1/2} \sum_{i=1}^n Y_i \mathbf{b}^T(\mathbf{x}_i) &= n^{1/2} n^{-1} \sum_{i=1}^n Y_i \mathbf{b}^T(\mathbf{x}_i) \mathbf{G}_n^{-1}(\boldsymbol{\theta}) \mathbf{G}_n(\boldsymbol{\theta}) = n^{1/2} \hat{\boldsymbol{\beta}}_n^T(\boldsymbol{\theta}) \mathbf{G}_n(\boldsymbol{\theta}); \\
n^{-1/2} \sum_{i=1}^n \mathbf{b}^T(\mathbf{x}_i) \hat{\boldsymbol{\beta}}_n(\boldsymbol{\theta}) \mathbf{b}^T(\mathbf{x}_i) &= n^{-1/2} \sum_{i=1}^n \hat{\boldsymbol{\beta}}_n^T(\boldsymbol{\theta}) \mathbf{b}(\mathbf{x}_i) \mathbf{b}^T(\mathbf{x}_i) = n^{1/2} \hat{\boldsymbol{\beta}}_n^T(\boldsymbol{\theta}) \mathbf{S}_n.
\end{aligned}$$

Thus for any  $\boldsymbol{\theta}$ ,

$$\begin{aligned}
\mathcal{S}_{nj}(\boldsymbol{\theta}) &= -\tilde{\lambda} n^{1/2} \{ \hat{\boldsymbol{\beta}}_n^T(\boldsymbol{\theta}) \mathbf{G}_n(\boldsymbol{\theta}) - \hat{\boldsymbol{\beta}}_n^T(\boldsymbol{\theta}) \mathbf{S}_n \} \mathbf{G}_n^{-1}(\boldsymbol{\theta}) \mathbf{R}_{j\theta}(\boldsymbol{\theta}) \hat{\boldsymbol{\beta}}_n(\boldsymbol{\theta}) \\
&= -\tilde{\lambda}^2 n^{1/2} \hat{\boldsymbol{\beta}}_n^T(\boldsymbol{\theta}) \mathbf{R}(\boldsymbol{\theta}) \mathbf{G}_n^{-1}(\boldsymbol{\theta}) \mathbf{R}_{j\theta}(\boldsymbol{\theta}) \hat{\boldsymbol{\beta}}_n(\boldsymbol{\theta}).
\end{aligned} \tag{A.6}$$

We let  $\theta_0$  be the limiting value of  $\hat{\boldsymbol{\theta}}$ . It is immediately clear from (A.6) that  $\theta_0$  solves

$$0 = \boldsymbol{\beta}_n^T(\theta_0) \mathbf{R}(\theta_0) \mathbf{G}_n^{-1}(\theta_0) \mathbf{R}_{j\theta}(\theta_0) \boldsymbol{\beta}_n(\theta_0). \tag{A.7}$$

Turning to the Hessian matrix, we see that by (A.4),

$$\begin{aligned}
n^{-1/2} \mathcal{M}_{n1,jk}(\boldsymbol{\theta}) &= -n^{-1} \sum_{i=1}^n \frac{\partial \hat{\boldsymbol{\beta}}_n^T(\boldsymbol{\theta})}{\partial \theta_j} \mathbf{b}(\mathbf{x}_i) \mathbf{b}^T(\mathbf{x}_i) \frac{\partial \hat{\boldsymbol{\beta}}_n(\boldsymbol{\theta})}{\partial \theta_k} \\
&= -n^{-1} \tilde{\lambda}^2 \sum_{i=1}^n \hat{\boldsymbol{\beta}}_n^T(\boldsymbol{\theta}) \mathbf{R}_{j\theta}^T(\boldsymbol{\theta}) \mathbf{G}_n^{-1}(\boldsymbol{\theta}) \mathbf{b}(\mathbf{x}_i) \mathbf{b}^T(\mathbf{x}_i) \mathbf{G}_n^{-1}(\boldsymbol{\theta}) \mathbf{R}_{k\theta}(\boldsymbol{\theta}) \hat{\boldsymbol{\beta}}_n(\boldsymbol{\theta}) \\
&= -\tilde{\lambda}^2 \hat{\boldsymbol{\beta}}_n^T(\boldsymbol{\theta}) \mathbf{R}_{j\theta}^T(\boldsymbol{\theta}) \mathbf{G}_n^{-1}(\boldsymbol{\theta}) \mathbf{S}_n^T \mathbf{G}_n^{-1}(\boldsymbol{\theta}) \mathbf{R}_{k\theta}(\boldsymbol{\theta}) \hat{\boldsymbol{\beta}}_n(\boldsymbol{\theta}).
\end{aligned}$$

Now using the fact that  $\widehat{\beta}_n(\boldsymbol{\theta}) = \beta_n(\boldsymbol{\theta}) + o_p(1)$  for any  $\boldsymbol{\theta}$ , we have at  $\boldsymbol{\theta}_0$  that

$$n^{-1/2}\mathcal{M}_{n1,jk}(\theta_0) = -\widetilde{\lambda}^2\beta_n^T(\theta_0)\mathbf{R}_{j\theta}^T(\theta_0)\mathbf{G}_n^{-1}(\theta_0)\mathbf{S}_n\mathbf{G}_n^{-1}(\theta_0)\mathbf{R}_{k\theta}(\theta_0)\beta_n(\theta_0) + o_p(1).$$

Similarly for the remaining term of the Hessian matrix we have

$$\begin{aligned} n^{-1/2}\mathcal{M}_{n2,jk}(\theta_0) &= \left[ n^{-1} \sum_{i=1}^n \{Y_i - \mathbf{b}^T(\mathbf{x}_i)\beta_n(\theta_0)\} \mathbf{b}^T(\mathbf{x}_i) \right] \frac{\partial^2 \beta_n(\theta_0)}{\partial \theta_{0j} \partial \theta_{0k}} + o_p(1) \\ &= n^{-1} \sum_{i=1}^n \epsilon(\mathbf{x}_i) \mathbf{b}^T(\mathbf{x}_i) \frac{\partial^2 \beta_n(\theta_0)}{\partial \theta_{0j} \partial \theta_{0k}} \\ &\quad + \left[ n^{-1} \sum_{i=1}^n \{g(\mathbf{x}_i) - \mathbf{b}^T(\mathbf{x}_i)\beta(\theta_0)\} \mathbf{b}^T(\mathbf{x}_i) \right] \frac{\partial^2 \beta_n(\theta_0)}{\partial \theta_{0j} \partial \theta_{0k}} + o_p(1) \end{aligned}$$

Now if we ignore the approximation error so that  $g(\mathbf{x}) \approx \mathbf{b}^T(\mathbf{x})\beta(\theta_0)$ , then

$$n^{-1/2}\mathcal{M}_{n2,jk}(\theta_0) = o_p(1).$$

Combining all the results, we now have that

$$n^{-1/2}\mathcal{M}_{n,jk}(\boldsymbol{\theta}_0) = -\widetilde{\lambda}^2\Lambda_{n,jk}(\boldsymbol{\theta}_0) + o_p(1), \quad (\text{A.8})$$

where  $\Lambda_{n,jk}(\boldsymbol{\theta}_0) = \beta_n^T(\theta_0)\mathbf{R}_{j\theta}^T(\theta_0)\mathbf{G}_n^{-1}(\theta_0)\mathbf{S}_n\mathbf{G}_n^{-1}(\theta_0)\mathbf{R}_{k\theta}(\theta_0)\beta_n(\theta_0)$ .

### A.3.2 Some Simple Asymptotic Theory

We of course define  $\widehat{\boldsymbol{\theta}}$  to solve  $\mathcal{S}_n(\boldsymbol{\theta}) = 0$ . Doing a Taylor series and assuming  $n^{1/2}$ -convergence of  $\widehat{\boldsymbol{\theta}}$ , we have that

$$0 = \mathcal{S}_n(\widehat{\boldsymbol{\theta}}) = \mathcal{S}_n(\theta_0) + n^{-1/2}\mathcal{M}_n(\theta_0)n^{1/2}(\widehat{\boldsymbol{\theta}} - \theta_0).$$

Let  $\boldsymbol{\Lambda}_n(\boldsymbol{\theta})$  denote the matrix with the  $(j, k)^{th}$  element as  $\Lambda_{n,jk}(\boldsymbol{\theta})$ . Hence using (A.8) we obtain

$$n^{1/2}(\widehat{\boldsymbol{\theta}} - \theta_0) = \widetilde{\lambda}^{-2}\boldsymbol{\Lambda}_n^{-1}(\theta_0)\mathcal{S}_n(\theta_0) + o_p(1). \quad (\text{A.9})$$

Now using (A.6) and (A.7), we see that

$$\begin{aligned}\mathcal{S}_{nj}(\theta_0) &= -\tilde{\lambda}^2 n^{1/2} \hat{\boldsymbol{\beta}}_n^T(\theta_0) \mathbf{R}(\theta_0) \mathbf{G}_n^{-1}(\theta_0) \mathbf{R}_{j\theta}(\theta_0) \hat{\boldsymbol{\beta}}_n(\theta_0) \\ &= -\tilde{\lambda}^2 n^{1/2} \{\hat{\boldsymbol{\beta}}_n(\theta_0) - \boldsymbol{\beta}_n(\theta_0)\}^T \mathbf{R}(\theta_0) \mathbf{G}_n^{-1}(\theta_0) \mathbf{R}_{j\theta}(\theta_0) \hat{\boldsymbol{\beta}}_n(\theta_0) \\ &\quad - \tilde{\lambda}^2 n^{1/2} \boldsymbol{\beta}_n^T(\theta_0) \mathbf{R}(\theta_0) \mathbf{G}_n^{-1}(\theta_0) \mathbf{R}_{j\theta}(\theta_0) \{\hat{\boldsymbol{\beta}}_n(\theta_0) - \boldsymbol{\beta}_n(\theta_0)\}.\end{aligned}$$

Define  $\mathcal{V}_j = \mathbf{R}(\theta_0) \mathbf{G}_n^{-1}(\theta_0) \mathbf{R}_{j\theta}(\theta_0)$  and  $\mathcal{W}_j = \mathcal{V}_j + \mathcal{V}_j^T$ . Then we have that

$$\mathcal{S}_{nj}(\theta_0) = -\tilde{\lambda}^2 \boldsymbol{\beta}_n^T(\theta_0) \mathcal{W}_j n^{1/2} \{\hat{\boldsymbol{\beta}}_n(\theta_0) - \boldsymbol{\beta}_n(\theta_0)\}. \quad (\text{A.10})$$

Now recall that  $\mathbf{S}_n \rightarrow \boldsymbol{\Omega}_1$  and  $\mathbf{G}_n(\theta_0) \rightarrow \boldsymbol{\Omega}_2(\theta_0)$  in probability. Hence we have that

$$\begin{aligned}n^{1/2} \{\hat{\boldsymbol{\beta}}_n(\theta_0) - \boldsymbol{\beta}_n(\theta_0)\} &= \mathbf{G}_n^{-1}(\theta_0) n^{-1/2} \sum_{i=1}^n \mathbf{b}(\mathbf{x}_i) \epsilon(\mathbf{x}_i) \\ &\rightarrow \text{Normal}\{0, \sigma_\epsilon^2 \boldsymbol{\Omega}_2^{-1}(\theta_0) \boldsymbol{\Omega}_1 \boldsymbol{\Omega}_2^{-1}(\theta_0)\},\end{aligned}$$

in distribution. So using (A.10) the  $(j, k)^{th}$  element of the covariance matrix of  $\mathcal{S}_n$  is given by

$$\text{cov}(\mathcal{S}_{nj}, \mathcal{S}_{nk}) = \tilde{\lambda}^4 \sigma_\epsilon^2 \boldsymbol{\beta}_n^T(\theta_0) \mathcal{W}_j \boldsymbol{\Omega}_2^{-1}(\theta_0) \boldsymbol{\Omega}_1 \boldsymbol{\Omega}_2^{-1}(\theta_0) \mathcal{W}_k \boldsymbol{\beta}_n(\theta_0) + o_p(1).$$

Hence using (A.9) we obtain

$$n^{1/2} \boldsymbol{\Sigma}_{n, \text{prop}}^{-1/2} (\hat{\boldsymbol{\theta}} - \theta_0) \rightarrow \text{Normal}(0, \mathbf{I}), \quad (\text{A.11})$$

where  $\boldsymbol{\Sigma}_{n, \text{prop}} = \boldsymbol{\Lambda}_n^{-1}(\theta_0) \mathcal{C}(\theta_0) \{\boldsymbol{\Lambda}_n^{-1}(\theta_0)\}^T$  with

$$\mathcal{C}_{jk}(\theta_0) = \sigma_\epsilon^2 \boldsymbol{\beta}_n^T(\theta_0) \mathcal{W}_j \boldsymbol{\Omega}_2^{-1}(\theta_0) \boldsymbol{\Omega}_1 \boldsymbol{\Omega}_2^{-1}(\theta_0) \mathcal{W}_k \boldsymbol{\beta}_n(\theta_0).$$



### A.3.3 Implementation

We propose the variance estimate as

$$\widehat{\Sigma}_{n,\text{prop}} = \mathbf{\Lambda}_n^{-1}(\widehat{\boldsymbol{\theta}}) \mathcal{C}(\widehat{\boldsymbol{\theta}}) \{\mathbf{\Lambda}_n^{-1}(\widehat{\boldsymbol{\theta}})\}^T, \quad (\text{A.12})$$

where  $\mathcal{C}(\widehat{\boldsymbol{\theta}})$  is a matrix whose  $(j, k)^{th}$  element is

$$\widehat{\mathcal{C}}_{jk} = n \widetilde{\lambda}^4 \widehat{\sigma}_\epsilon^2 \widehat{\boldsymbol{\beta}}^T(\widehat{\boldsymbol{\theta}}) \widehat{\mathcal{W}}_j \mathbf{G}_n^{-1}(\widehat{\boldsymbol{\theta}}) \mathbf{S}_n \mathbf{G}_n^{-1}(\widehat{\boldsymbol{\theta}}) \widehat{\mathcal{W}}_k \widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\theta}})$$

and  $\mathbf{\Lambda}_n(\widehat{\boldsymbol{\theta}}) = \sum_{i=1}^n \partial \Psi_i(\widehat{\boldsymbol{\theta}}) / \partial \boldsymbol{\theta}^T$ . Here  $\widehat{\sigma}_\epsilon^2$  is the estimated variance of  $\epsilon(\mathbf{x}_i)$  and can be calculated by first fitting a standard spline regression and then forming the residual variance. Also,  $\widehat{\mathcal{W}}_j = \widehat{\mathcal{V}}_j + \widehat{\mathcal{V}}_j^T$ , where  $\widehat{\mathcal{V}}_j = \mathbf{R}(\widehat{\boldsymbol{\theta}}) \mathbf{G}_n^{-1}(\widehat{\boldsymbol{\theta}}) \mathbf{R}_{j\theta}(\widehat{\boldsymbol{\theta}})$ .

The above estimator requires analytic expression of  $\partial \widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$  and  $\partial \Psi_i(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ . These quantities could be obtained using the implicit function theorem, which is introduced as follows. Dependence on  $\boldsymbol{\theta}$  is dropped where appropriate.

To find the first-order derivative of  $\widehat{\boldsymbol{\beta}}$  with respect to  $\boldsymbol{\theta}$ , take total derivative with respect to  $\boldsymbol{\theta}$  on both sides of the identity  $\partial J(\boldsymbol{\beta} | \boldsymbol{\theta}) / \partial \boldsymbol{\beta} |_{\widehat{\boldsymbol{\beta}}} = 0$ , we get

$$\frac{d}{d\boldsymbol{\theta}} \left( \frac{\partial J}{\partial \boldsymbol{\beta}} \Big|_{\widehat{\boldsymbol{\beta}}} \right) = \frac{\partial^2 J}{\partial \boldsymbol{\theta}^T \partial \boldsymbol{\beta}} \Big|_{\widehat{\boldsymbol{\beta}}} + \frac{\partial^2 J}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\beta}} \Big|_{\widehat{\boldsymbol{\beta}}} \frac{\partial \widehat{\boldsymbol{\beta}}}{\partial \boldsymbol{\theta}} = 0.$$

Assuming that  $\partial^2 J / \partial \boldsymbol{\beta}^T \partial \boldsymbol{\beta} |_{\widehat{\boldsymbol{\beta}}}$  is non-singular, which is true for our model, we obtain the analytic expression of the first-order derivative of  $\widehat{\boldsymbol{\beta}}$  as,

$$\frac{\partial \widehat{\boldsymbol{\beta}}}{\partial \boldsymbol{\theta}} = - \left( \frac{\partial^2 J}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\beta}} \Big|_{\widehat{\boldsymbol{\beta}}} \right)^{-1} \left( \frac{\partial^2 J}{\partial \boldsymbol{\theta}^T \partial \boldsymbol{\beta}} \Big|_{\widehat{\boldsymbol{\beta}}} \right). \quad (\text{A.13})$$

It is easily seen from (3.9) that

$$\frac{\partial^2 J}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\beta}} = 2\{\mathbf{B}^T \mathbf{B} + \lambda \mathbf{R}(\boldsymbol{\theta})\}, \quad (\text{A.14})$$

and that

$$\frac{\partial^2 J}{\partial \boldsymbol{\theta}^T \partial \boldsymbol{\beta}} = 2\lambda \frac{\partial}{\partial \boldsymbol{\theta}} \{ \mathbf{R}(\boldsymbol{\theta}) \boldsymbol{\beta} \}. \quad (\text{A.15})$$

Substitute the above into (A.13) and we have

$$\frac{\partial \hat{\boldsymbol{\beta}}}{\partial \boldsymbol{\theta}} = -\lambda \{ \mathbf{B}^T \mathbf{B} + \lambda \mathbf{R}(\boldsymbol{\theta}) \}^{-1} \left[ \frac{\partial}{\partial \boldsymbol{\theta}} \{ \mathbf{R}(\boldsymbol{\theta}) \boldsymbol{\beta} \} \Big|_{\hat{\boldsymbol{\beta}}} \right]. \quad (\text{A.16})$$

The first-order derivative of  $\Psi_i(\boldsymbol{\theta})$  with respect to  $\boldsymbol{\theta}$  is, for  $i = 1, \dots, n$ ,

$$\frac{\partial \Psi_i}{\partial \boldsymbol{\theta}} = \sum_{k=1}^K b_k(\mathbf{x}_i) \{ Y_i - \mathbf{b}^T(\mathbf{x}_i) \hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) \} \frac{\partial^2 \hat{\boldsymbol{\beta}}_k}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} - \left( \frac{\partial \hat{\boldsymbol{\beta}}}{\partial \boldsymbol{\theta}} \right)^T \mathbf{b}(\mathbf{x}_i) \mathbf{b}^T(\mathbf{x}_i) \left( \frac{\partial \hat{\boldsymbol{\beta}}}{\partial \boldsymbol{\theta}} \right).$$

To find the second-order derivative of  $\hat{\boldsymbol{\beta}}_k$  with respect to  $\boldsymbol{\theta}$ , take the second-order total derivative with respect to  $\boldsymbol{\theta}$  on both sides of the identity  $\partial J / \partial \boldsymbol{\beta}_k|_{\hat{\boldsymbol{\beta}}_k} = 0$ , we get, for  $k = 1, \dots, K$ ,

$$\begin{aligned} \frac{d^2}{d\boldsymbol{\theta}^T d\boldsymbol{\theta}} \left( \frac{\partial J}{\partial \boldsymbol{\beta}_k} \Big|_{\hat{\boldsymbol{\beta}}_k} \right) &= \frac{d}{d\boldsymbol{\theta}^T} \left\{ \frac{d}{d\boldsymbol{\theta}} \left( \frac{\partial J}{\partial \boldsymbol{\beta}_k} \Big|_{\hat{\boldsymbol{\beta}}_k} \right) \right\} \\ &= \frac{\partial^3 J}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T \partial \boldsymbol{\beta}_k} \Big|_{\hat{\boldsymbol{\beta}}_k} + \frac{\partial^3 J}{\partial \boldsymbol{\theta} \partial \boldsymbol{\beta}_k^2} \Big|_{\hat{\boldsymbol{\beta}}_k} \frac{\partial \hat{\boldsymbol{\beta}}_k}{\partial \boldsymbol{\theta}^T} + \frac{\partial^2 J}{\partial \boldsymbol{\beta}_k^2} \Big|_{\hat{\boldsymbol{\beta}}_k} \frac{\partial^2 \hat{\boldsymbol{\beta}}_k}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \\ &\quad + \frac{\partial^3 J}{\partial \boldsymbol{\beta}_k^3} \Big|_{\hat{\boldsymbol{\beta}}_k} \frac{\partial \hat{\boldsymbol{\beta}}_k}{\partial \boldsymbol{\theta}} \frac{\partial \hat{\boldsymbol{\beta}}_k}{\partial \boldsymbol{\theta}^T} = \mathbf{0}. \end{aligned}$$

Obviously for our method, we have that  $\partial^3 J / \partial \boldsymbol{\beta}_k^3 \equiv 0$ , so the last term in the above result disappears. Assuming that  $\partial^2 J / \partial \boldsymbol{\beta}_k^2|_{\hat{\boldsymbol{\beta}}_k} \neq 0$ , then the analytic expression for the second-order derivative of  $\hat{\boldsymbol{\beta}}_k$  is obtained as,

$$\frac{\partial^2 \hat{\boldsymbol{\beta}}_k}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = - \left( \frac{\partial^2 J}{\partial \boldsymbol{\beta}_k^2} \Big|_{\hat{\boldsymbol{\beta}}_k} \right)^{-1} \left( \frac{\partial^3 J}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T \partial \boldsymbol{\beta}_k} \Big|_{\hat{\boldsymbol{\beta}}_k} + \frac{\partial^3 J}{\partial \boldsymbol{\theta} \partial \boldsymbol{\beta}_k^2} \Big|_{\hat{\boldsymbol{\beta}}_k} \frac{\partial \hat{\boldsymbol{\beta}}_k}{\partial \boldsymbol{\theta}^T} \right). \quad (\text{A.17})$$

To complete the calculation, we need to know the following quantities,

$$\frac{\partial \mathbf{R}(\boldsymbol{\theta}) \boldsymbol{\beta}}{\partial \boldsymbol{\theta}}, \quad \frac{\partial^3 J}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T \partial \boldsymbol{\beta}_k}, \quad \text{and} \quad \frac{\partial^3 J}{\partial \boldsymbol{\theta} \partial \boldsymbol{\beta}_k^2},$$

all of which involve derivatives of  $\mathbf{R}(\boldsymbol{\theta})$  with respect to  $\boldsymbol{\theta}$ . The ensuing derivation depends on the particular PDE model of interest.

#### A.3.4 Implementation with PDE Example (3.2)

We explain in this section the calculation of above model dependent quantities, in the context of the PDE model (3.2). We know

$$\mathbf{f}(\mathbf{x}; \boldsymbol{\theta}) = \frac{\partial \mathbf{b}(\mathbf{x})}{\partial t} - \theta_D \frac{\partial^2 \mathbf{b}(\mathbf{x})}{\partial z^2} - \theta_S \frac{\partial \mathbf{b}(\mathbf{x})}{\partial z} - \theta_A \mathbf{b}(\mathbf{x}),$$

where  $\mathbf{b}(\mathbf{x}) = \{b_1(\mathbf{x}), \dots, b_K(\mathbf{x})\}^T$  is the vector of basis functions. The matrix  $\mathbf{R}(\boldsymbol{\theta})$  is shown in (A.1). In this example, the coefficients of matrices  $\mathcal{B}_\ell$ 's are

$$\begin{aligned} r_1(\boldsymbol{\theta}) &= \theta_D^2, & r_2(\boldsymbol{\theta}) &= \theta_S^2, & r_3(\boldsymbol{\theta}) &= \theta_A^2, & r_4(\boldsymbol{\theta}) &= \theta_D \theta_S, & r_5(\boldsymbol{\theta}) &= \theta_D \theta_A, \\ r_6(\boldsymbol{\theta}) &= \theta_S \theta_A, & r_7(\boldsymbol{\theta}) &= -\theta_D, & r_8(\boldsymbol{\theta}) &= -\theta_S, & r_9(\boldsymbol{\theta}) &= -\theta_A, & r_{10}(\boldsymbol{\theta}) &= 1. \end{aligned}$$

Then we have

$$\begin{aligned} \frac{\partial \mathbf{R}(\boldsymbol{\theta}) \boldsymbol{\beta}}{\partial \boldsymbol{\theta}} &= \frac{\partial}{\partial \boldsymbol{\theta}} \sum_{\ell=1}^L r_\ell(\boldsymbol{\theta}) \mathcal{B}_\ell \boldsymbol{\beta} \\ &= \sum_{\ell=1}^L \mathcal{B}_\ell \boldsymbol{\beta} \frac{\partial r_\ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}. \end{aligned}$$

Notice that  $(\partial^2 J / \partial \boldsymbol{\theta} \partial \boldsymbol{\beta}_k)^T$  is the  $k^{th}$  row of  $\partial^2 J / \partial \boldsymbol{\theta}^T \partial \boldsymbol{\beta}$  given in (A.15). Let  $\tilde{\mathbf{b}}_{\ell,k}$  be the  $k^{th}$  row of  $\mathcal{B}_\ell$ , then  $\mathcal{B}_\ell = (\tilde{\mathbf{b}}_{\ell,1}^T, \dots, \tilde{\mathbf{b}}_{\ell,K}^T)^T$ . Then, we could write

$$\frac{\partial^2 J}{\partial \boldsymbol{\theta} \partial \boldsymbol{\beta}_k} = 2\lambda \sum_{\ell=1}^L \tilde{\mathbf{b}}_{\ell,k} \boldsymbol{\beta} \frac{\partial r_\ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}.$$

Then

$$\frac{\partial^3 J}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T \partial \boldsymbol{\beta}_k} = 2\lambda \sum_{\ell=1}^L \tilde{\mathbf{b}}_{\ell,k} \boldsymbol{\beta} \frac{\partial^2 r_\ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}. \quad (\text{A.18})$$

In this simulated example

$$\frac{\partial^2 r_1(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \frac{\partial^2 r_2(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \frac{\partial^2 r_3(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 2 \end{pmatrix},$$

and  $\frac{\partial^2 r_\ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \equiv \mathbf{0}$ , for  $\ell = 4, \dots, 10$ . Notice that  $\partial^2 J / \partial \boldsymbol{\beta}_k^2$  is the  $k^{th}$  diagonal element of  $\partial^2 J / \partial \boldsymbol{\beta}^T \partial \boldsymbol{\beta}$  given in (A.14), then

$$\frac{\partial^3 J}{\partial \boldsymbol{\theta} \partial \boldsymbol{\beta}_k^2} = 2\lambda \sum_{\ell=1}^L \mathcal{B}_\ell(k, k) \frac{\partial r_\ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \quad (\text{A.19})$$

where  $\mathcal{B}_\ell(k, k)$  is the  $k^{th}$  diagonal element of the matrix  $\mathcal{B}_\ell$ .

Finally, substituting  $\partial^2 J / \partial \boldsymbol{\beta}_k^2$ , (A.18) and (A.19) into (A.17) results in the expression of  $\partial^2 \hat{\boldsymbol{\beta}}_k / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T$ . The matrices  $\mathcal{B}_1, \dots, \mathcal{B}_{10}$  are calculated using Simpson's rule, see to **Supplemental Material** Appendix A.2 for detailed calculation.

#### A.4 Full Conditional Distributions for Bayesian Method

To sample from the posterior distribution (3.15) using Gibbs sampler, we need full conditional distributions of all the unknowns. Due to conjugacy, parameters  $\sigma_\epsilon^2$  and  $\gamma_\ell$ 's have close form full conditionals. Define  $\text{SSE} = (\mathbf{Y} - \mathbf{B}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{B}\boldsymbol{\beta})$ . If we define "rest" to mean conditional on everything else, we have

$$\begin{aligned} [\sigma_\epsilon^2 | \text{rest}] &\propto (\sigma_\epsilon^2)^{-(a_\epsilon + n/2) - 1} \exp\{-(b_\epsilon + \text{SSE}/2)/\sigma_\epsilon^2\} \\ &= \text{IG}(a_\epsilon + n/2, b_\epsilon + \text{SSE}/2), \\ [\gamma_0 | \text{rest}] &\propto \gamma_0^{a_0 + K/2 - 1} \exp\{-b_0 \gamma_0 - \gamma_0 \boldsymbol{\zeta}^T(\boldsymbol{\beta}, \boldsymbol{\theta}) \boldsymbol{\zeta}(\boldsymbol{\beta}, \boldsymbol{\theta})/2\} \\ &= \text{Gamma}(a_0 + K/2, b_0 + \boldsymbol{\zeta}^T(\boldsymbol{\beta}, \boldsymbol{\theta}) \boldsymbol{\zeta}(\boldsymbol{\beta}, \boldsymbol{\theta})/2), \\ [\gamma_1 | \text{rest}] &\propto \gamma_1^{a_1 + K/2 - 1} \exp\{-b_1 \gamma_1 - \boldsymbol{\beta}^T(\gamma_1 H_1 + \gamma_1 \gamma_2 H_3) \boldsymbol{\beta}/2\} \\ &= \text{Gamma}(a_1 + K/2, b_1 + \boldsymbol{\beta}^T(H_1 + \gamma_2 H_3) \boldsymbol{\beta}/2), \\ [\gamma_2 | \text{rest}] &\propto \gamma_2^{a_2 + K/2 - 1} \exp\{-b_2 \gamma_2 - \boldsymbol{\beta}^T(\gamma_2 H_2 + \gamma_1 \gamma_2 H_3) \boldsymbol{\beta}/2\} \\ &= \text{Gamma}(a_2 + K/2, b_2 + \boldsymbol{\beta}^T(H_2 + \gamma_1 H_3) \boldsymbol{\beta}/2). \end{aligned}$$

The parameters  $\boldsymbol{\beta}$  and  $\boldsymbol{\theta}$  do not have closed form full conditionals, which are instead

$$\begin{aligned} [\boldsymbol{\beta}|\text{rest}] &\propto \exp\{-\boldsymbol{\beta}^T(\sigma_\epsilon^{-2}\mathbf{B}^T\mathbf{B} + \gamma_1 H_1 + \gamma_2 H_2 + \gamma_1 \gamma_2 H_3)\boldsymbol{\beta}/2 \\ &\quad - \sigma_\epsilon^{-2}\boldsymbol{\beta}^T\mathbf{B}^T\mathbf{Y} - \gamma_0 \boldsymbol{\zeta}^T(\boldsymbol{\beta}, \boldsymbol{\theta})\boldsymbol{\zeta}(\boldsymbol{\beta}, \boldsymbol{\theta})/2\}, \\ [\boldsymbol{\theta}|\text{rest}] &\propto \exp\{-\boldsymbol{\theta}^T\boldsymbol{\theta}/(2\sigma_\theta^2) - \gamma_0 \boldsymbol{\zeta}^T(\boldsymbol{\beta}, \boldsymbol{\theta})\boldsymbol{\zeta}(\boldsymbol{\beta}, \boldsymbol{\theta})/2\}. \end{aligned}$$

To draw samples from these full conditionals, a Metropolis-Hastings update within the Gibbs sampler is applied. In the special case of a linear PDE, the model error is also linear in  $\boldsymbol{\beta}$ , represented by  $\boldsymbol{\zeta}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \mathbf{F}(\boldsymbol{\theta})\boldsymbol{\beta}$ . Then the term  $\boldsymbol{\zeta}^T(\boldsymbol{\beta}, \boldsymbol{\theta})\boldsymbol{\zeta}(\boldsymbol{\beta}, \boldsymbol{\theta})$  is a quadratic function in  $\boldsymbol{\beta}$ . Define  $\mathbf{H} = \mathbf{H}(\boldsymbol{\theta}) = \gamma_0 \mathbf{F}^T(\boldsymbol{\theta})\mathbf{F}(\boldsymbol{\theta}) + \gamma_1 H_1 + \gamma_2 H_2 + \gamma_1 \gamma_2 H_3$ , and  $\mathbf{D} = \{\mathbf{B}^T\mathbf{B} + \sigma_\epsilon^2 \mathbf{H}(\boldsymbol{\theta})\}^{-1}$ . By completing the square in  $[\boldsymbol{\beta}|\text{rest}]$ , the full conditional of  $\boldsymbol{\beta}$  under linear PDE models is

$$\begin{aligned} [\boldsymbol{\beta}|\text{rest}] &\propto \exp[-(2\sigma_\epsilon^2)^{-1}\{\boldsymbol{\beta}^T(\mathbf{B}^T\mathbf{B} + \sigma_\epsilon^2 \mathbf{H})\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\mathbf{B}^T\mathbf{Y}\}] \\ &= \text{Normal}(\mathbf{D}\mathbf{B}^T\mathbf{Y}, \sigma_\epsilon^2 \mathbf{D}). \end{aligned}$$

## VITA

Xiaolei Xun received her B.S. in Statistics in July 2007 from Zhejiang University, China. She entered the Department of Statistics at Texas A&M University in August 2007 and received her M.S. in Statistics in December 2009, and her Ph.D. in Statistics in May 2012 from Texas A&M University, College Station. Her research interests include Bayesian methodology, statistical inverse problems, and functional data analysis. She is a student member of American Statistical Association (ASA) and International Society of Bayesian Analysis (ISBA). She could be reached at xiaolei.xun@gmail.com. Her address is: Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843.

The typist for this dissertation was Xiaolei Xun.